UNIVERSITAT DE BARCELONA

Neural network mechanisms of working memory interference

João Moura Barbosa

UNIVERSITAT DE BARCELONA

Programa de Doctorat en Biomedicina

Area: Neuroscience - Neurophysiology and computation in cortical systems

Neural network mechanisms of working memory interference

João Moura Barbosa

Supervised by:
Albert Compte Braquets

Barcelona, 2019

“A vida acontece entre dois cagares”

António Barbosa (1925-2019)

Table of contents

Acknowledgements

Long before I started filling this empty page, parts of it surfaced in my mind. Countless people, directly or indirectly, were actors in the play whose script I finally wrote. I am hereby acknowledging some of those people. Warning: occasionally, but intentionally, I will confound acknowledgment with applause or even with lame sentimental ventilation.

I will start by thanking the people outside the lab - back when I had a life outside the lab: hopefully one day you will forgive my long periods of absence and welcome me again in your privileged circles. First, *La Famiglia*, my partners in crime for which I never found a replacement in all the places I have been. *Alex, Amandi, Black, Couto, Manel, Neto, Trevis e Tomás, obrigado por aturarem o meu crescente mas ainda amador cepticismo, muitas vezes realmente, eu sei, um insuportável serdocontrismo.*

Aproveito também para agradecer aos meus pais. Pelas implícitas razões que leva um filho a agradecer aos seus pais, mas também em particular à Mãe, por nunca teres desistido de me convencer que só vale a pena fazer o que está, verdadeiramente, correcto. Ao Pai, por cedo me teres motivado a questionar tudo e todos, sem piedade. Ambas, vejam só, são características essenciais em quem aspira ser um cientista que se preze. Também quero agradecer ao resto da família, à Mariana, ao Manel e à Adelaide e, também, aos 4 avós, os legítimos embaixadores da família do Marco e da Foz. Obrigado por serem a âncora que sempre me recordará de onde venho. Frustra-me que as oportunidades de demonstrar o amor e admiração que nutro por vocês - família e Famiglia - sejam tão escassas. Mas, sempre que encontrar uma, nem que seja escondida numa tese que ninguém vai ler, eu vou aproveitar. Gostaria tanto, muito mais que qualquer outra coisa ao meu alcance, que a idade nunca nos pudesse separar. Aproveito o lanço em português para agradecer aos Zucas, Miguel e Bernardo: agora que que acabei a porra da tese espero conseguir juntá-los em muitos mais churrascos dos que fizemos até hoje.

También debo expresar mi eterno cariño por Cata, por haberme dado más que yo logré retribuir y por me haber permitido el honor de hacer de sus amigos los míos: Majo, Pietro, Miguel y varios mas, muchas gracias.

Then, those within the lab. I want to acknowledge my deepest gratitude and admiration for Albert Compte, a true mentor: I have to thank you for giving me all the freedom to explore and expand, always fuelling the illusion that I was driving on my own. But, most importantly, for showing me and everyone around you that only the humble can achieve true wisdom. I also would like to acknowledge how lucky I was to meet a person with such a rare intellectual drive as Jaime de la Rocha. You are a true rockstar, very much like the greatest de la Rocha*, Zack de la Rocha*.

Then *Zorra*, for filtering out all those stupid, random ideas and transforming all the mediocre ideas into really great ones. I truly hope our friendship will evolve in real

life, when both leave the lab. Ah, also thank you for *living with me*. Which brings me to acknowledge Heike's omnipresence: your *questioning eyes* already had a strong impact on who I am. I also need to acknowledge Ainhoa, *la jefa*, for being the only one patient enough to argue with me for that long and for all those years - how many years exactly? To all the people in the lab today - Lucia, Lejla, Dani Linares, Diego, David el Borde, David el Mago, Adrià, Pablo, Yerko, Dani Duque, Molano, Alba y Balma or Balma y Alba, Tiffany, Alex H., Lluis, Jordi - and yesterday - the greatest of all, the *Chilean Man*, Dani Jercog, Maira, Klaus: for showing me that it is possible to take it easy and to fucking nail it both in science and in life - and countless others in the BARCCSYN community - Ramon, partner of loud concerts, Josefina, *maestra de la vida*, Iñigo, Maria Alemany, Gabriela, Phillip, Txema, all the Neurochats community - and so many other people that I can only hope they know my admiration for you. Thank you, in one way or another, you helped me grow.

I also would like to thank the world-class neuroscientists who accepted to be in my thesis defense committee, and so become the only 3 people who will read past this page (hopefully): Alex Roxin, Athena Akrami, Mark Stokes, thank you.

This has been (not finished yet!) a long ride, which I went through at very different speeds. During the first months of this *trip*, I was focused on pretending I was working much more than I actually was. Sometimes, even struggling to cover up the strongest hangover. Unexpectedly, I ended up embarrassed of spending so much time in the lab, and pretended I was working much less than I actually was.

Obrigado, this was great fun!

1. INTRODUCTION

1.1 What is working memory?

Our ability to memorize is at the core of our cognitive abilities. How could we effectively make decisions without considering memories of previous experiences? Broadly, our memories can be divided in two categories: long-term and short-term memories. Sometimes, short-term memory is also called working memory and throughout this thesis I may use both terms interchangeably. As the names suggest, long-term memory is the memory you use when you remember concepts for a long period, such as your name or age, while short-term memory is the system you engage while choosing between different wines at the liquor store. As your attention jumps from one bottle to another, you need to hold in memory characteristics of previous ones to pick your favourite. By the time you pick your favourite bottle, you might remember the prices or grape types of the other bottles, but you are likely to forget all of those details an hour later at home, opening the wine in front of your guests[1]. The computer is a metaphor often used to explain our current knowledge about how the brain works, and memory is likely to be the most intuitive part of that metaphor[2]. Much like humans, the computer also has long and short-term memory systems - the hard drive and the Random Access Memory (RAM). While I am writing this document, temporary changes are kept with high fidelity in the RAM of my laptop but, unless I commit those changes to my hard-drive, those will be lost forever upon reboot. At least in abstract terms, this might seem how human memory works. In addition to obvious differences in the hardware implementation (living cells versus transistors), the actual features are also very different. One obvious difference is how we control the flow between those two types of memories. As in my example above, committing a new paragraph to long-term memory is as easy as to click “save”. How we transfer a transient memory of a cell phone number into long-term memory is much more complicated than that and remains a mystery. For the purpose of this thesis, I will ignore that mystery and instead focus exclusively on visual short-term memory. Another, perhaps not as intuitive difference between a computer’s and the brain’s short-term memory is its fidelity. Until my laptop's RAM is full or I shut it down,

[1] Surely after you drink it you will forget all those details, but for reasons not covered in this thesis.
[2] This is not a coincidence, since the modern computer was in fact modeled after the human mind. Alan Turing’s groundbreaking definition of a “Turing Machine”, the first model of a modern computer, was actually Turing’s attempt to model the human mind at his early twenties. Turing’s impact in modern society is monumental and immeasurable. Moreover, because he was the first to see the human mind as a computing machine, he is arguably the pioneer of both Cognitive Sciences and Artificial Intelligence (Copeland et al. 2017).

every word I wrote will be kept virtually forever with high fidelity. That is not the case with our short-term memory. We cannot remember a cell phone number for much longer than a couple of minutes - perhaps only seconds - without the help of any kind of long-term memory, such as our own or a piece of paper. In addition, in contrast with the computer's, different memories might interfere with each other. Can you imagine the words of this document interfering with each other as I am writing them? The broad aim of this thesis is to study the neural mechanisms underlying visual working memory interference. Before I expose my own findings, I will try to provide a concise but hopefully thorough review of previous findings from which my own surfaced.

1.2 Behavioral evidence for working memory limitations

How to study working memory

Inspired by Jean Piaget's classical experiments[3], early experiments with monkeys probing their working memory involved showing them two plates, one of them having an edible reward. During a given delay period, both plates were covered and hidden from the monkey's sight, arguably engaging the monkey's working memory system. After a delay period of a few seconds, the monkey was allowed to chose any plate. Figure 1.1a illustrates this experiment.

Nowadays, visual working memory is studied by presenting a stimulus with the specific feature of interest, say a location (Figure 1.1b, sample), followed by a blank screen during a given period (Figure 1.1b, delay) and finally asking the subject to recall that feature from their memory (Figure 1.1b, report). This experimental design is also called delayed-estimation. For example, Figure 1.1b illustrates a typical experimental design for those interested in studying visual working memory of

[3] In Piaget's experiment, a child is shown two boxes, one of them containing a toy. The boxes are then closed and, after a brief delay, the child is asked to select which box contains the toy. After several correct answers, the experimenter switches the toy's location on the children's sight. The experiment continues to test how easily the child changes its response from the previous location to the new one. This task is considered to be the earliest method to test working memory (Lowe et al. 2009) and has been shown to correlate with degree of maturity of the subject's prefrontal cortex (Diamond and Goldman-Rakic 1989).

locations - delayed-estimation of location. In experiments with monkeys, such as the macaque, the report usually involves saccading[4] to the recalled location, so eye position is typically monitored using eye-trackers or eye implants[5]. In addition, in cases where brain activity is simultaneously recorded, it is recommended to constrain eye movements to the center of the screen (marked by the fixation point) during the sample and delay period. This ensures the recorded brain activity reflects stimuli or memories that have a constant reference point and are not contaminated by other visual stimuli outside of the experimental control. To probe working memory limits, several parameters of this experimental design can be modulated, including the delay duration and the number of to-be-remembered objects - the set size (e.g one location vs three colors, Figure 1.1b,c).

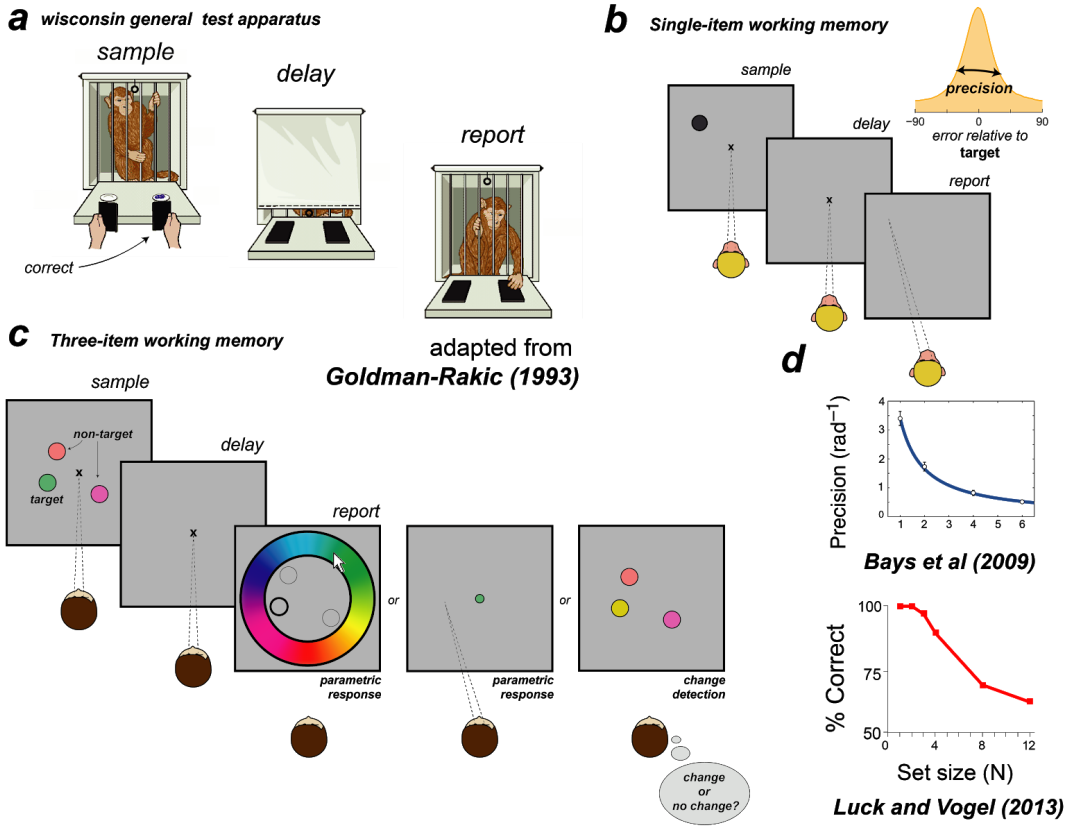


Figure 1.1. How working memory is studied in the lab. **a)** Schematic of a monkey performing the delayed-response task in the WGTA. This task, an adaptation of Piaget's original experiments for monkeys, was used in (Fuster and Alexander 1971). **b)** The oculomotor delayed-response task (ODR task), a modification by (Funahashi et al. 1989) of a task

[4] A saccade is a quick eye movement, in contrast with smooth pursuit movements, where the eyes move smoothly.
[5] An eye tracker is a device that measures eye positions. This tracking can be done by means of an implant or in more modern approaches by processing video images of the eye.

originally introduced by (Hikosaka and Wurtz 1983). In this task, the monkey is required to maintain gazing at the central fixation dot while a to-be-memorized location is cued at 1 out of 8 different locations. After a delay period of varying length, the monkey is allowed to make saccades that are rewarded at the cued location. **c)** Three examples of multi-item working memory tasks. During the sample period, three colored dots are presented, which the subject has to remember. After a delay period, the figure illustrates three alternative designs: i) the target item is revealed by cueing its location and the subject has to parametrically report its color on a color wheel ii) the target item is revealed by cueing its color and the subject has to parametrically report its location iii) a set of items is presented again and the subject has to report if they are different from the original ones. **d)** (Top) parametric response paradigms reveal a power law decrease in precision with increasing set sizes, while (bottom) change detection paradigms point to a plateau in performance until the maximum capacity is reached (3-4 items), and a decay to chance level after crossing it.

Working memory fidelity decays with delay duration

With a study published in German and later re-published in English by (Laming and Laming 1992), Friedrich Hegelmaier was perhaps the first scientist to study the relationship of working memory decay with increasing delay durations. In his experiment, Hegelmaier measured human working memory precision for line lengths. Countless others studies have replicated Hegelmaier's finding for different visual features (Phillips and Baddeley 1971; Zhang and Luck 2009; Barrouillet et al. 2012; McKeown and Mercer 2012; Pertzov et al. 2017; Pertzov et al. 2013; Shin et al. 2017; Bliss et al. 2017). However, some have either found modest or no significant decay of working memory fidelity with increasing delays, in particular studies using delayed match to sample[6] tasks of spatial frequency (e.g. (Greenlee et al. 1993)) or speed (e.g. (Greenlee et al. 1995)), but also in delayed estimations of motion direction and coherence (Blake et al. 1997). In spite of these intriguing null findings, there is growing evidence that, at least for visuo-spatial working memory (i.e. working memory for locations, central to this thesis), longer delay durations impact memory fidelity. These different dynamics of forgetting have been used as an argument supporting separate storage processes for different stimulus features (Pasternak and Greenlee 2005). Figure 1.2 shows how delay duration increases saccades' spread around each target for a monkey performing a visuo-spatial working memory task with 8 locations.

[6] Delayed match-to-sample tasks, a general case of the change detection task introduced in Figure 1c, demand binary responses such as "change/no change", "higher/lower", "clockwise/counterclockwise" etc.

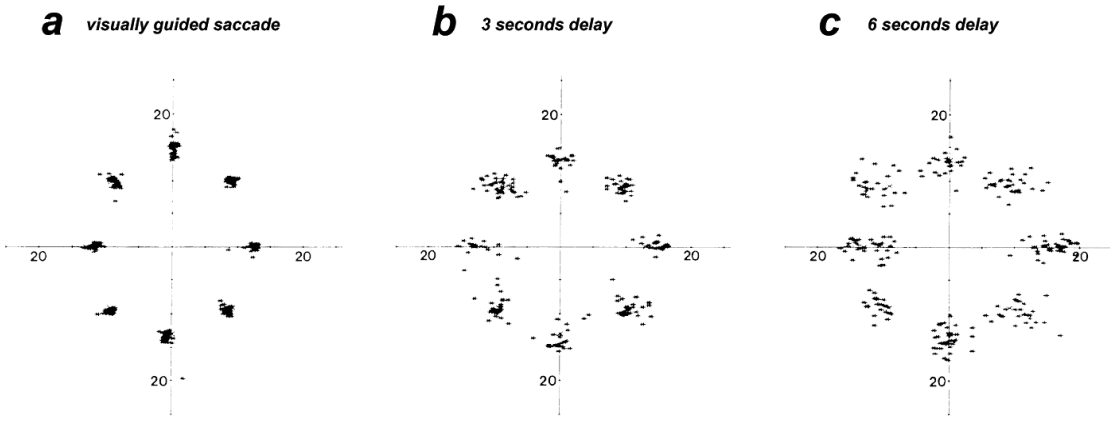


Figure 1.2. Decrease in precision with increasing delay length. Figure taken from (Funahashi et al. 1989)**.**

Working memory capacity

Although not the first working memory feature to be characterized (see Working memory fidelity decays with delay duration, above), capacity is arguably the most interesting feature of any memory system. In humans, working memory capacity has been found to correlate with overall cognitive ability (Shipstead et al. 2016). The finding that human working memory - then called primary memory - is of limited capacity dates back to (James 1890), but its actual capacity was not determined until much later. In what came to be a controversial article under the title “The magical number seven, plus or minus two”, (Miller 1956) claimed to have found the capacity of human working memory - regarded until today as Miller’s law. The very first sentence of his article is the following:

> *“My problem is that I have been persecuted by an integer. For seven years this number has followed me around, has intruded in my most private data, and has assaulted me from the pages of our most public journals.”*

While this was the paragraph that closed the article:

> *“And finally, what about the magical number seven? What about the seven wonders of the world, the seven seas, the seven deadly sins, the seven daughters of Atlas in the Pleiades, the seven (...) For the present I propose to withhold judgment.*

Perhaps there is something deep and profound behind all these sevens, something just calling out for us to discover it. But I suspect that it is only a pernicious, Pythagorean coincidence."

In his famous article[7] Miller revises previously published data from different working memory experiments, showing the surely entertained reader that the number 7 was, in fact, everywhere[8].

Using multi-item working memory change detection - a modern experimental design where human subjects have to detect changes in stimulus features in a set of varying size (Figure 1.1c, change detection) - (Luck and Vogel 1997) found that human working memory had in fact a much lower capacity of between 3 and 4 items. The evidence for such numbers was found by counting change detection hits (correct responses) during experimental trials of varying set sizes. Crucially, Luck and Vogel found that the fraction of correct trials was maximal (100%) for set sizes of up to 3-4 items, but dropped abruptly with set sizes larger than that (Figure 1.1d, bottom). This finding gave rise to the slot model, which predicts that until all the 3-4 "slots", constituting our memory capacity are full, memory resolution should be held constant for all items. (Cowan 2001) later reviewed previous studies and published a meta-analysis under the mandatory title "The magical number 4 in short-term memory".

With a simple, but key modification of the original multi-item task, (Wilken and Ma 2004) introduced an experimental design called delayed-estimation. In this design, subjects' reports are no longer binary (i.e. change/no change). Instead, similarly to what had been the standard in single-item working memory, subjects reported by making an eye movement to, or a mouse click in, the remembered probed location (Figure 1.1b, parametric response). Plotting precision[9] (instead of percentage of correct responses) against increasing set sizes revealed that precision did not drop abruptly, but smoothly, following a power law. In other words, the resolution with which items are kept in memory depends on the set size, even for sizes below the

[7] Up to 29106 citations when I wrote these lines
[8] Hidden in what superficially seems to be a numerology account of working memory, there was a much relevant, revolutionary finding. In this article, Miller describes how humans could increase their effective storage capacity through the use of item grouping ("chunking").
[9] Precision is defined as the inverse of the variability around the correct report. The less variable, the more precise.

assumed capacity. This finding was important as it is incompatible with the slot model, i.e. a fixed-item working memory capacity. Rather, it seemed that different items held in working memory share common resources - the so-called resource model.

Taking advantage of subjects' parametric responses, (Zhang and Luck 2008) established another cornerstone of working memory capacity research. Instead of measuring precision directly from responses around the target (Figure 1.1b, precision), they fitted a mixture model of a Gaussian distribution around the target and a uniform distribution that accounts for random guessing (see Methods for a detailed formulation). When using the standard approach, the authors replicated a smooth decrease in precision with set size. However, when accounting for random guessing they revealed that working memory precision reached a plateau (a critical feature of the slot-model; but see Binding of independent features in working memory, below). In a more recent study, (Adam et al. 2017) modified the classical experiment and pushed the slot model hypothesis a bit further. In her experiment, human subjects reported the color of all items (varying on each trial between 1-4 and 6) in a randomly defined sequential order. Their hallmark finding was that after the 3rd or 4th item, subjects' response distributions were essentially random, pointing to a fixed limit of working memory capacity - compatible with the slot model but not with the resource model.

Interference between different memories

Studying working memory capacity requires increasing set sizes, which reveals yet other working memory limitations. Simultaneous items, it has been shown, interact with each other depending on factors such as delay duration (Shin et al. 2017) or feature similarity (Emrich and Ferber 2012; Almeida et al. 2015; Nassar et al. 2018). Moreover, not only simultaneously but also previously memorized items interfere with current memories (Kiyonaga et al. 2017). In this section, I will introduce two such sources of interference: swap errors and serial biases. These interferences are central to my thesis, considering that I i) helped to uncover a novel interference between simultaneous items (*Chapter 4.1*), ii) found the neural correlates of such interferences *(Chapter 4.2, 4.3)* and iii) propose neural mechanisms underlying both swap *(Chapter 4.1)* and serial errors *(Chapter 4.2)* by implementing two computational models that reproduce i) and ii).

Binding of independent features in working memory

How the conjunctions of different visual features are kept in mind is a long standing question in cognitive neuroscience - the so-called binding problem. In the same study that argued for human working memory capacity to lie between 3 to 4 items, (Luck and Vogel 1997) found that memorizing two independent features - in this case, color and orientation - was as difficult as to memorize a conjunction of the two. The idea of increasing memory capacity through “chunking” dates back to (Miller 1956) (see Working memory capacity, above), but that binding came without cost, suggesting that complex items were stored as a whole, not through independent storage of each of their features. Several subsequent studies, however, contradicted this hypothesis (Delvenne and Bruyer 2004; Olson and Jiang 2002; Parra et al. 2011; Xu 2002; Wheeler and Treisman 2002). Amongst them, evidence was found that different visual features decay with different timescales, as already mentioned above (*Fidelity decays with delay duration*). Moreover, another prediction of storing complex objects as a whole, instead of their independent features, is that errors when reporting each feature should be correlated: if one object’s memory is corrupted, all of its features should reflect it. Two studies from different labs (Fougnie and Alvarez 2011; Bays, Wu, et al. 2011) found, instead, that errors for color and orientation were essentially independent when remembering complex objects with both features, yet again supporting the idea of independent storage systems.

Another piece of evidence in favour of the independent storage of features is a phenomenon called swap errors. Hidden in what Zhang and Luck modeled as guesses, (Bays et al. 2009) found accurate reports relative to non-target stimuli (see Figure 1.3a,b, swaps). Using the same mixture model approach as (Zhang and Luck 2008), (Bays et al. 2009) quantified those swap errors by including an additional Gaussian component in the original mixture model fit, which, for each trial, is centered at non-target stimuli (see Methods). When plotting precision relative to targets against set size, similarly to what was done by (Zhang and Luck 2008), but excluding both guesses and swap errors, the authors found, once again, that precision decayed smoothly with set size, following a power law - new evidence for the resource model and against the slot model. Follow-up studies have shown that swap errors increase with delay duration, set-size and feature similarity (Figure 1.3). A modern, non-parametric method (Bays 2016) revealed that swap errors can occur

in as much as 40% of the trials with a set size of 8 items. Experimental designs that require subjects to rate their confidence on a trial-by-trial basis show, however, that swap errors occur more often in low-confidence trials, suggesting that at least some proportion of swap errors might be in fact smart guesses (Mitchell et al. 2018; Pratte 2018), an interpretation supported as well by findings of (Adam et al. 2017) described in Working memory capacity.

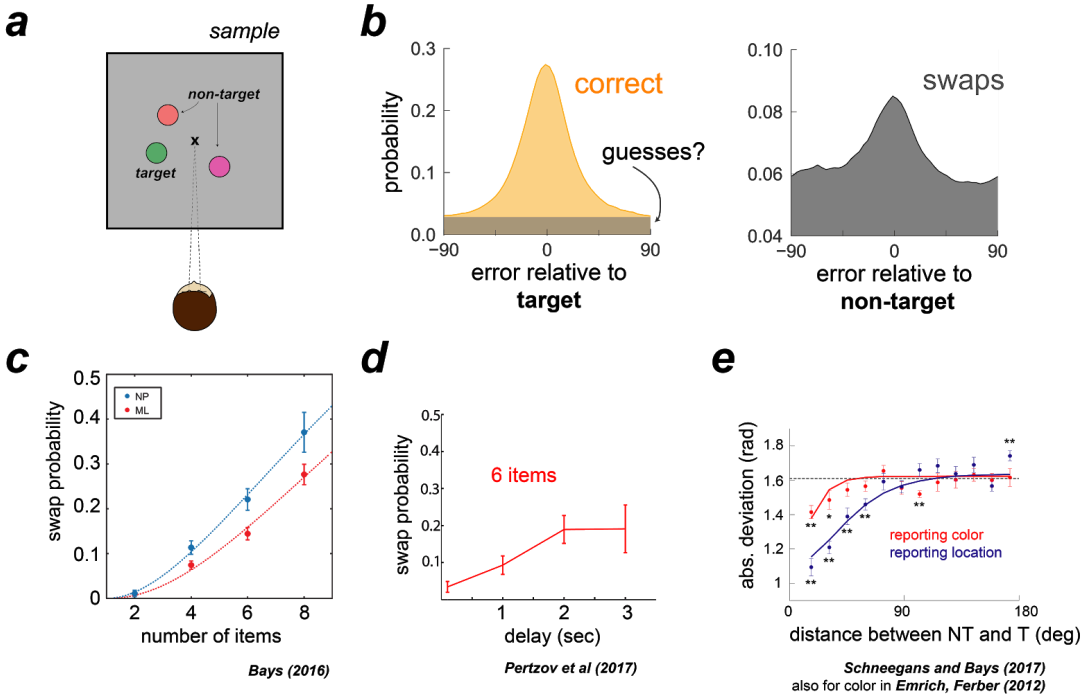


Figure 1.3. Swap errors are modulated with several working memory tasks parameters. **a)** Repetition of Figure 1.1c. In each trial of multi-item working memory trials, without the subject's knowledge at the time of presentation, one item will be probed and is thus called the target while the other, non-probed items are called non-targets. **b)** Data from a study I conducted on-line: on the left, error distribution relative to target, and on the right, error distribution computed relative to non-targets. Hidden in what seem to be guessing errors (in gray on the left), there are trials - called swap errors - where subjects actually respond correctly relative to non-targets (on the right). **c)** Swap probability estimated with mixture model fit (red) and with a non-parametric method (blue). Both methods reveal that swap errors increase with set-size and **d)** delay duration. **e)** Absolute report distance to non-targets is smaller when they are closer to targets (in the non-probed feature space), meaning that subjects make more swap errors when the non-targets' non-probed feature is similar to targets'.

Interference from previously memorized items

Two studies - almost simultaneously - found that previous memories interfere with forthcoming stimuli, in ways that current trial reports are biased towards previous stimuli. This so-called serial dependence or serial bias has been described

experimentally using many different paradigms (Kiyonaga et al. 2017; Bliss et al. 2017; Xia et al. 2015; Manassi et al. 2018; Czoschke et al. 2018; Alais et al. 2018; Manassi et al. 2017; Samaha et al. 2018; Suárez-Pinilla et al. 2018; Fischer and Whitney 2014; Liberman et al. 2016; Alexi et al. 2018; Cicchini et al. 2014; Fritsche et al. 2017; Taubert, Alais, et al. 2016; Taubert, Van der Burg, et al. 2016). In particular, serial dependence was observed in paradigms requiring the delayed estimation of visual features (Kiyonaga et al. 2017), such as orientation (Fischer and Whitney 2014; Fritsche et al. 2017; Samaha et al. 2018), numerosity (Cicchini et al. 2017), location (Bliss et al. 2017; Papadimitriou et al. 2017; Papadimitriou et al. 2015), facial identity (Liberman et al. 2014) or body size (Alexi et al. 2018). In this thesis we also demonstrate serial dependence in color working memory (*Chapter 4.2*). The curve of serial biase is obtained by plotting the current trial error as a function of the distance between the target in current and previous trial. Figure 1.4b illustrates the common pattern of serial biases for orientation working memory, while Figure 1.4c shows a similar profile (however weaker in magnitude) for location. A characteristic feature of serial dependence is that attractive biases are stronger when previous and current stimuli are similar. In some cases, also repulsive biases have been found for farther distances (Figure 1.4c), (Fritsche et al. 2017; Samaha et al. 2018).

It has been speculated that these ubiquitous attractive biases are a consequence of the world's tendency to be stable, and that biases have the functional role of averaging internal noise (Cicchini et al. 2018; Kiyonaga et al. 2017; Fischer and Whitney 2014; Cicchini et al. 2014). Some have further argued that serial dependence is of adaptive nature, changing its strength depending on the stimulus statistics (Cicchini et al. 2018; Kiyonaga et al. 2017; Fischer and Whitney 2014; Taubert, Alais, et al. 2016; Cicchini et al. 2014).

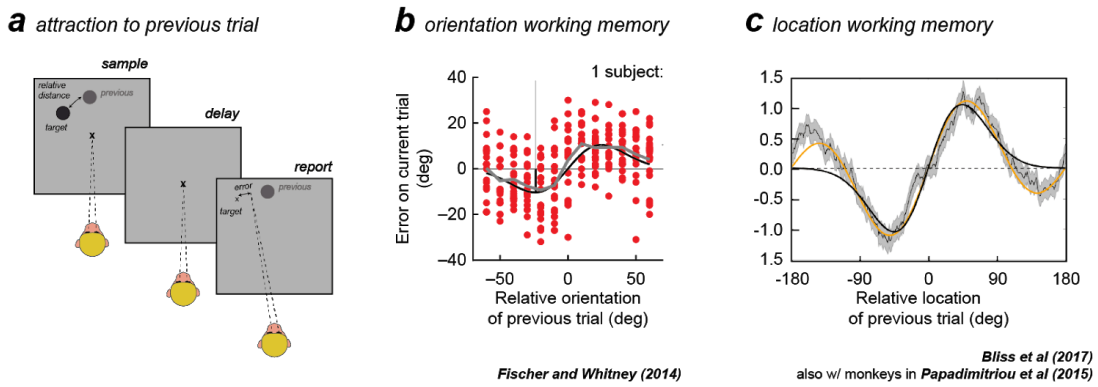


Figure 1.4. Serial dependence in working memory tasks. **a)** Illustration of a trial with attraction to the previous stimulus. The final report is slightly attracted to the previous stimulus, the location of which was overlaid for illustration. **b)** Plotting the average error against relative orientations - or locations in **c)** - of the previous trial reveals attractive biases (positive) for trials in which previous stimuli were close to the current stimulus and repulsive biases (negative) when they were far.

1.3 Neural basis of working memory

A neuron and its receptive field

The human brain is a network of ~1×10^{10} neurons, interacting through ~1×10^{14} connections, called synapses. To provide some perspective, approximations say that the number of stars in the Milky Way is roughly ~1×10^{10}, the same number as neurons in our brain. At any time point, each neuron integrates the input of all neurons it is connected with. If this global input sums higher than a certain threshold, that neuron itself will send a binary output to all neurons that are, in turn, connected to its output terminals. The neurons that receive this binary signal can be different neurons than the ones which helped our neuron to reach its threshold. If that is the case, this sub-network is said to be feedforward. Otherwise, if some neuron feeds its output back to the neurons whose signal it integrates, these connections are recurrent, or feedback connections.

The brain can be divided into subregions, each of which can be seen as a network of neurons with its internal intricate connectivity patterns. Moreover, some of these networks rest close to the sensory organs, listening directly to the outside world, while others lie in deep structures of the brain that only receive input from other brain

structures. Each neuron's output is then a result of a complex interaction with all neurons connected to it through feedforward and feedback connections. This complicated maze of connections results in neural selectivity to features of the scene of increasing abstractness, depending on how deep in the brain's hierarchy a neuron is situated. For example, the visual receptive field of a neuron in early visual cortex, two synapses away from the retina, is a particular region of the visual space in which a stimulus would modify its activity[10]. Neurons in early visual cortex are mostly selective to simple objects, such as oriented lines, while neurons in orbitofrontal cortex (OFC), a region very high up in the brain hierarchy, can represent complex concepts such as confidence (Kepecs et al. 2008).

Persistent activity as the neural correlate of working memory

Until the 1950's, neurobiologists often denied that neuronal mechanisms of working memory - or any other high order cognitive abilities - could ever be unveiled (Goldman-Rakic 1993). Nevertheless, the finding that lesioning certain brain regions such as prefrontal cortex (PFC) impaired working memory (Warren et al. 1957), paved the way for the groundbreaking work of (Fuster and Alexander 1971) and (Kubota and Niki 1971). Almost simultaneously, these two groups - in the USA and Japan, respectively - found what is until today regarded by many as the neural correlate of working memory. Both studies trained monkeys to perform variations of the Wisconsin General Test Apparatus (Figure 1.1a), while recording from their PFC neurons using fine electrodes previously introduced into their brains. Both studies found that during the delay period, when monkeys had to remember which plate to choose (Figure 1.1a, delay period), some neurons kept firing throughout the whole interval. In retrospective, while both studies were equally remarkable, (Fuster and Alexander 1971) were more explicit envisioning neuronal sustained firing as the neural correlates of short-term memory. On the other hand, (Kubota and Niki 1971) related their findings to delayed motor execution, relative to the decision where to move made before the delay. In fact, both studies were confounding working memory with motor execution - neuronal delay activity could be reflecting monkeys' movement. This confound was solved in Patricia S. Goldman-Rakic lab, by carefully controlling movements during the mnemonic period of a oculomotor working memory

[10] David H. Hubel and Torsten Wiesel won the Nobel Prize in Physiology or Medicine in 1981, for their seminal work on information processing in the visual system. Experimenting with cats, they described how signals from the eye are processed in the cortex and generate the building blocks of the visual scene: neurons that act as edge detectors, motion detectors, stereoscopic depth detectors and color detectors.

task. In this seminal work, (Funahashi et al. 1989) trained two monkeys to perform a parametric working memory task (Figure 1.1b, see also *How to study working memory* and *Working memory fidelity decays with delay duration*) and recorded from their PFC neurons. Once again, neurons with delay activity were found in PFC, but this time, the authors also found selectivity to the memorized location (Figure 1.5), without the motor execution confound.

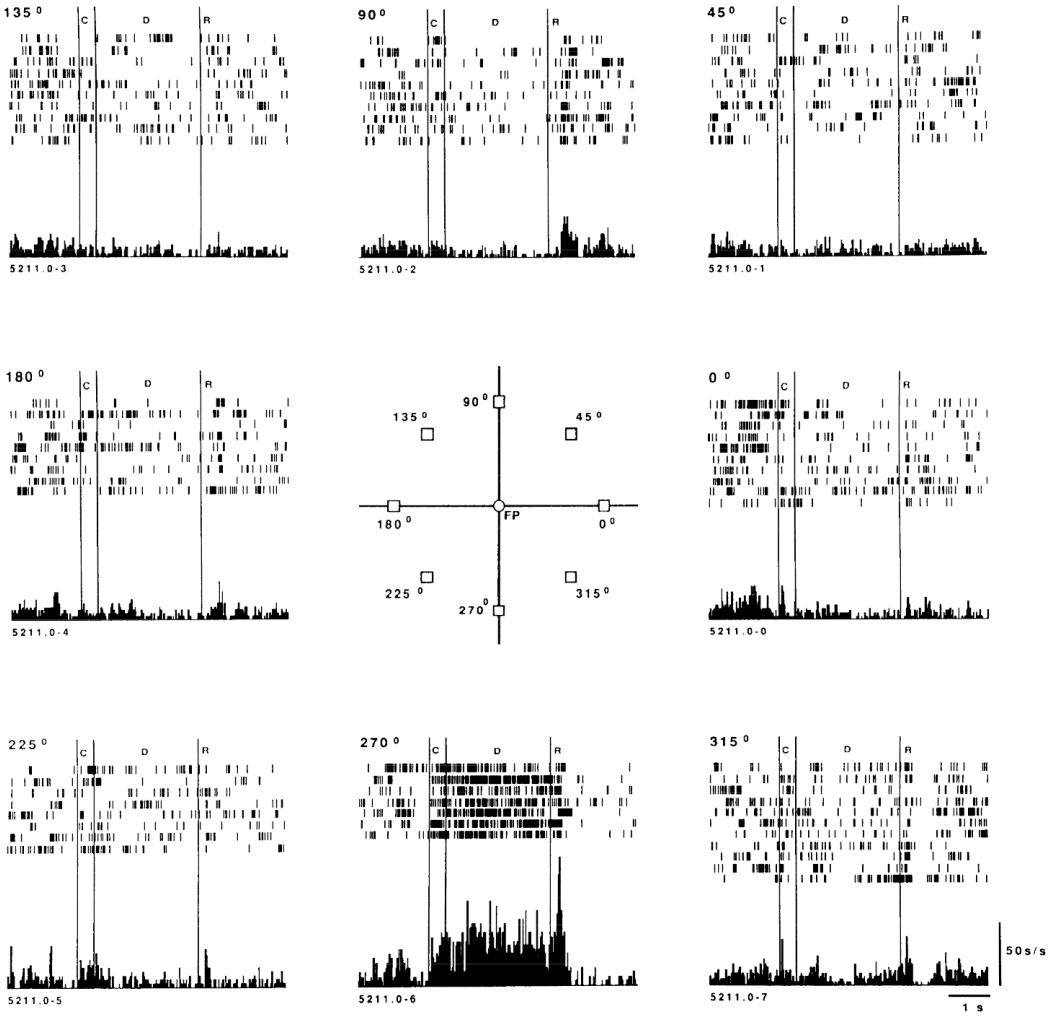


Figure 1.5. An example neuron recorded from prefrontal cortex (PFC) during a visuospatial working memory task. Each of the 8 panels shows neural activity of the same neuron during several trials. Each panel corresponds to different locations the monkey was memorizing in different trials. First two vertical lines represent the time when the monkey was visually cued for the to-be-memorized location, while the last vertical is the time the monkey was allowed to saccad. In between, there is a delay period of 3 seconds, when the monkey had to remember the cued location. The neuron depicted in this figure shows elevated activity during the delay period when memorizing the bottom center location. Reproduced from (Funahashi et al. 1989).

Similar to receptive fields, but in the absence of any external stimulus, (Funahashi et al. 1989) described for the first time the existence of mnemonic fields - regions of the visual scene which, when memorized by the monkey, would modify the activity of a specific neuron in PFC. Figure 1.5 illustrates this phenomenon very clearly for one neuron. Much like for neurons in sensory areas such as V1[11], if one plots averaged neuronal activity against different stimulus parameters, a bell-shaped tuning curve is revealed. This sustained tuning during mnemonic periods has been successfully modeled using attractor networks[12] (Compte et al. 2000; Wimmer et al. 2014; Wang 1999; Durstewitz et al. 2000) *(see Neural network models of working memory, Methods).*

Silent and dynamic code and other challenges to the stable code hypothesis[13]

Despite the fact that some neurons are persistently active during the maintenance period in working memory tasks, most recorded neurons are not (Zaksas and Pasternak 2006; Barak et al. 2010; Jun et al. 2010). This has led to the emergence of alternative hypotheses concerning the neural bases of working memory. For example, instead of relying on a stable code accomplished by persistent activity in single neurons, information could be stored "silently" in enhanced synaptic strengths through short-term plasticity - the synaptic hypothesis; (Mongillo et al. 2008). Alternatively, a stable representation might be achieved by combining transient activity of a large population of neurons, each of which active at different epochs of the delay period - the dynamic code hypothesis; (Goldman 2009; Druckmann and Chklovskii 2012; Zaksas and Pasternak 2006). It is important to stress that in both hypotheses, single neurons are not persistently active and their selectivity is not stable during the delay. However, only models based on the synaptic hypothesis require synaptic plasticity mechanisms, also called 'activity-silent' mechanisms. More details are given in the section *Network models of working memory, Methods*. While theoretically appealing, and despite doubtless existence of short-term plasticity in the brain (Wang et al. 2006), there was virtually no neurophysiological evidence for its role in working memory maintenance. Recently, indirect evidence has surfaced from neuroimaging studies that I describe below, after introducing the fundamental use of information

[11] Visual cortex, located on the back of the brain, is divided into several sub-regions. The most important regions are V1, V2, V3 and V4. Their numbers reflect how high they are up in the visual hierarchy.
[12] Attractor networks are dynamical networks endowed with strong recurrence. These are called "attractor" because their dynamics evolve towards, i.e. are attracted to, a stable pattern over time.
[13] This section includes parts of a larger comment (Barbosa 2017) that I published in the Journal of Neuroscience called "Working Memories Are Maintained in a Stable Code".

decoders (for technical details see *Decoding stimulus information, Methods*) to discard different models from neuroimaging datasets.

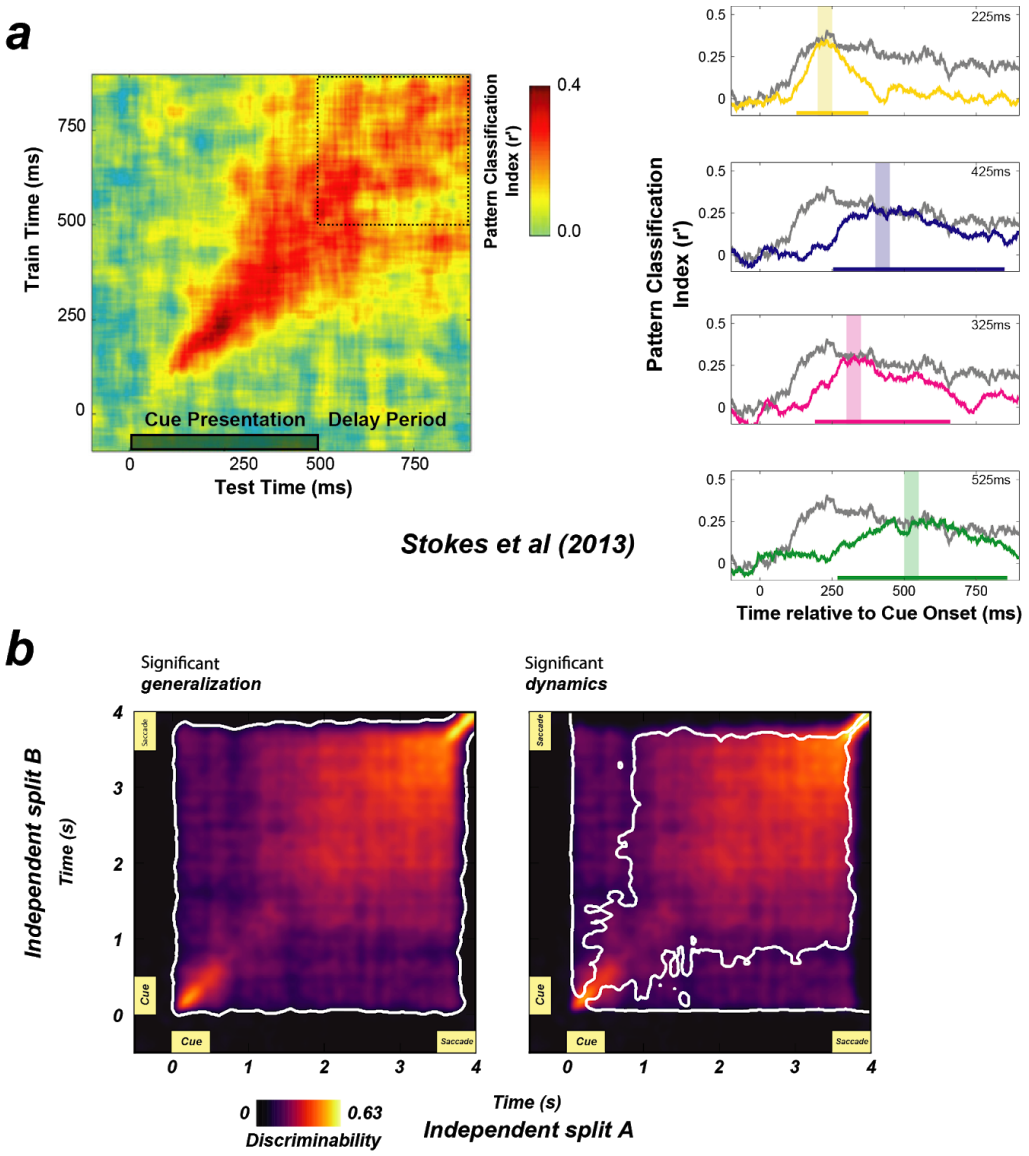


Figure 1.6. Cross-temporal decoders reveal stable dynamics during the delay period of working memory tasks. **a)** Left, a cross-temporal decoder from a task with a short delay period. Dashed line, added to the original figure, marks the delay period where the decoding accuracy is maximal for all time point combinations, suggesting the code to be stable for this period. This method is problematic because, without further analyses, it relies on subjective interpretations of color temperature. A better approach on the right, where decoders were trained at different time points (from cue presentation to early-delay, red to green) and tested throughout the whole trial period. Solid colored bars on the bottom of each panel mark time points for which the stimulus was decoded significantly above chance. The delay-decoder, in green, shows the most stable code. **b)** Cross-temporal decoder from another study where the authors directly compared a stable decoder with a dynamic decoder. On the left, white line marks time points for which the decoder is significantly better than chance. On the right, the

line marks periods in which a stable decoder performs worse than the dynamic decoder. It can be seen that by mid-delay the system is stabilized and the code remains stable until the trial is finished. Dynamic codes during cue presentation have been found extensively and their underlying neural mechanisms are discussed in (Barbosa 2017).

To assess whether the same neural code is used during different trial periods - a prediction of the stable code hypothesis - it is common to train a decoder on neural activity during one period of a task and use this decoder to extract stimulus information from neural activity recorded during other periods - here called "cross-temporal analysis," as in the study by (Stokes 2015). If the decoder is able to extract similar amounts of stimulus information from different trial periods, the code is said to be stable or generalizable across those periods (Figure 1.6). Alternatively, if information at a given time point can only be extracted by a decoder trained at that same time point, the code is said to be dynamic. On the other hand, support of the synaptic hypothesis frequently comes from null results: decoders of neuroimaging data that fail to extract stimulus information in a certain period of the trial are interpreted to fail because of the absence of such information in the neuronal activity, rather than for reasons related to the limitations of the methods that were employed. In fact, when using a particular decoding method on a particular neural signal, a decoding failure can happen for other reasons than absence of information. As we developed in *Chapter 4.3*, It can be that a particular brain region activity was still correlated with the absent stimulus, but the decoder was not able to detect it because i) the noise inherent to the recording method was higher than a potentially weak signal or ii) the decoding method was inappropriate for the specific kind of signal. Moreover, when information becomes available at a later trial period, that code is often said to be reactivated - arguably from a hidden, *'activity-silent'* synaptic trace. Again, this interpretation neglects potential limitations associated with neuroimaging techniques. For example, decoding methods applied to EEG[14] recordings are known to consider strongly parieto-occipital electrodes (Foster et al. 2016; Foxe et al. 1998; Reinhart et al. 2012). Therefore, a failure to decode using such methods tells more about occipital lobe rather than about the whole brain. Before we go in depth into which areas are involved (frontal vs sensory areas), I will reinterpret neuroimaging studies that claim evidence for the synaptic hypothesis.

[14] Electroencephalography (EEG) is an non-invasive method that records the electrical activity of the brain. Typically, it consists of placing electrodes on the scalp to measure voltage fluctuations within the brain. It is characterized by high temporal resolution (<1 msec), but poor spatial resolution - typically tens of electrodes.

An example of such is a recent fMRI[15] study by (Sprague et al. 2016), in apparent contradiction with *'activity-based'* working memory maintenance. In that study, a retro cue (Griffin and Nobre 2003) - introduced during the delay of a two-item working memory task - selectively increased information about the uncued feature of the cued item. This was interpreted as information recovery from a hidden, synaptic trace. However, (Watanabe and Funahashi 2014), in a data set later reanalyzed by (Spaak et al. 2017), found a similar effect that supports a different interpretation. In a dual task, with an attention task encapsulated within the delay of a working memory task, selectivity of prefrontal neurons increased after completing the attention task. However, because spiking activity did not stop carrying stimulus information through the working memory delay, this suggests that synaptic working memory is unnecessary to explain the fMRI findings by Sprague et al. (2016). Indeed, in a task similar to that used in the study by (Sprague et al. 2016), (Cisek and Kalaska 2005) recorded single units from the monkey premotor cortex and also found that information about a memorized location increased after retro-cueing it. In line with the findings of (Sprague et al. 2016), neurons keeping information about the then irrelevant location decreased their activity, suggesting that neural populations holding different locations in working memory might inhibit each other. In fact, (Cisek 2006) later modeled this information increase through a decrease in mutual inhibition between two neural populations without taking any plasticity mechanism into account. While reactivation from a synaptic trace is still a possible cause of the information increase reported in the studies by (Spaak et al. 2017; Watanabe and Funahashi 2014) and (Sprague et al. 2016), there is an alternative explanation under the framework of competing neural populations (Cisek 2006). More recently, (Schneegans and Bays 2017b) modeled BOLD responses using this framework and replicated all main findings of Sprague et al (2016) without including synaptic plasticity in their model.

Two independent studies (Rose et al. 2016; Wolff et al. 2017) reported that if an item previously stored in working memory was outside the focus of attention during the delay, modern decoding methods failed to detect stimulus information in BOLD or EEG signals. When the memory was brought back to the focus of attention, a significant amount of information was again detectable using the same methods.

[15] Functional magnetic resonance imaging (fMRI captures the blood-oxygenation-level (or BOLD) signals, which measures the oxygenation/deoxygenation of blood cells in the brain, as an indirect measure of local neural activity. It is a widely used non-invasive technique to measure brain activity. It is characterized by a high spatial resolution (3x3 mm cubes) but a poor temporal resolution (~ 1 sec).

These findings were interpreted as evidence in favor of the synaptic hypothesis. Critically, one limitation of these studies is the lack of direct access to single-unit activity; instead, they rely on signals that are an average of large neural populations (Dubois et al. 2015), orders of magnitude larger than the selectivity clusters found in cortical areas associated with working memory - for example, posterior parietal cortex (PPC) and prefrontal cortex (PFC) (Masse et al. 2017). By analyzing single-unit activity in a conceptually similar task, (Watanabe and Funahashi 2014) overcame this limitation and provided evidence in favor of an alternative interpretation. Specifically, to solve a dual task, monkeys had to memorize a cued location while attending somewhere else. Much like in studies by (Rose et al. 2016; Wolff et al. 2017), when the memory was outside the focus of attention, neurons exhibited a decrease in stimulus selectivity. Nevertheless, even under the most difficult attention conditions, persistent activity was still significantly above baseline through the delay and carried stimulus information. This point can be strengthened by results of the cross-temporal decoding analysis on spiking activity in (Spaak et al. 2017): as in the other two data sets, the mnemonic code was generalizable and remained stable throughout the delay period when monkeys were actively attending elsewhere. Because information was never absent from spiking activity in PFC, synaptic working memory mechanisms once again appear unnecessary. More recently, using a very similar task, (Christophel et al. 2018) confirmed that lack of decodability from previous studies was, in fact, due to lack of power. By increasing the number of subjects dramatically (n=87), the authors could decode both the attended and unattended memory item from BOLD signals coming from IPS and FEF, areas that are downstream from the occipital cortex.

The main findings of those studies (Rose et al. 2016; Wolff et al. 2017), are in fact more challenging to the stable code hypothesis. (Rose et al. 2016) found that immediately after the stimulus information became unaccessible to their decoder, a single, strong TMS pulse recovered that information as if the memory was reactivated from a hidden trace (Figure 1.7). (Wolff et al. 2017) finding is related, but more limited, claiming that a non-specific visual stimulus increased memory information measured by their decoder. Importantly, decoded memory information never really dropped to chance, which is reminiscent of the finding by (Sprague et al. 2016), where cueing a memory increased that memory information.

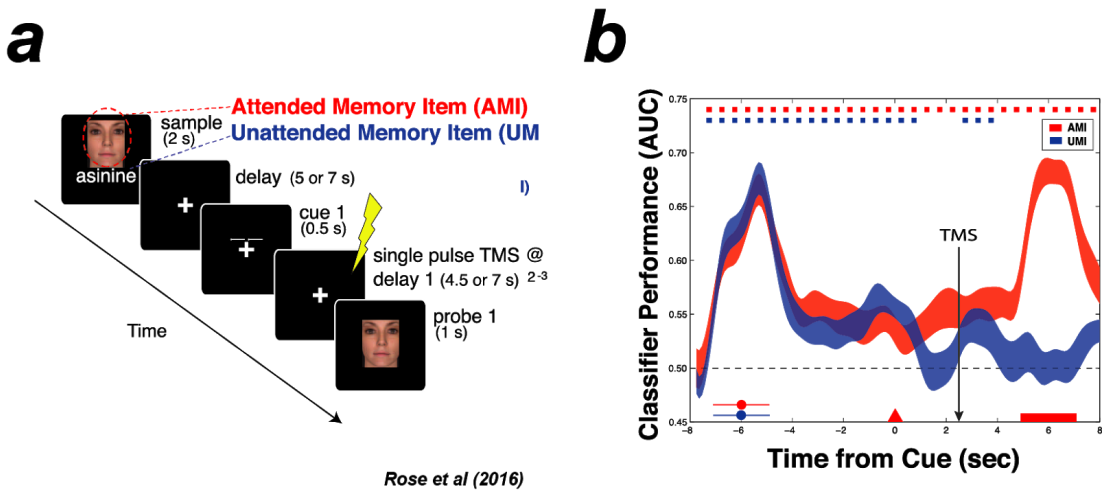


Figure 1.7. Unspecific TMS[16] impulse reactivates lost working memory representations. **a)** A two-item working memory task which starts by presenting two stimuli - a face and a word - followed by a delay during which both stimuli remain relevant, until one of the items is cued - on this example trial, the face. A second delay follows, after which a second stimulus is probed (match/non-match). Omitted from the original figure for compactness is the rest of the trial, see the original publication for further details. **b)** Classification accuracy for each of the two stimuli, categorized prospectively as attended (AMI, red) and unattended (UMI, blue) by the upcoming cue at 0 sec (word and face, respectively in example trial shown in a). Before cueing, classification accuracies are similar for both attended and unattended items. After cueing, the information about the unattended stimulus drops to chance but, critically, its information can be recovered by means of a non-specific TMS pulse.

Finally, (Lundqvist et al. 2016; Lundqvist et al. 2018) have further challenged the stable code hypothesis, a rare case of electrophysiological evidence for the synaptic hypothesis. Their reasoning is based on similar assumptions as the synaptic theory hypothesis from (Mongillo et al. 2008) which is described in detail in *Network models of working memory, Methods*. Briefly, reanalyzing data sets previously interpreted as evidence for persistent activity (Lundqvist et al. 2016; Lundqvist et al. 2018) realized that what was seen as persistent activity could be a trial-averaging artifact. Indeed, it could be that single neurons are active at random moments and, when averaging the activity of many neurons over many trials, one could be misled by an artificially flat, stable code. They tested their hypothesis by detecting the timings of activity bursts during the delay period of each trial, and found that during encoding and delay period, there was a higher burst rate which also increased with working memory load (Figure 1.10c). However, it is not clear if this finding is inconsistent with a stable code with additional, sporadic bursts of activity. In fact, the method used to detect burst

[16] Transcranial magnetic stimulation (TMS) is a non-invasive stimulation technique, in which a magnetic field is targeted at a small brain region to cause electric current to flow through the neural tissue.

consists of detecting timings where activity was 3 standard deviations above the mean which, depending on the distribution family, will eventually happen on some small fraction of the trials. Additionally, as crucially pointed out by (Constantinidis et al. 2018; Li et al. 2018), this hypothesis predicts a much lower variability during baseline compared to the delay period, when the stochastic bursts supposedly occur more often. This prediction has been invalidated by two independent labs (Chang et al. 2012; Qi and Constantinidis 2012). Finally, recent theory-driven intracellular recordings experiments (Inagaki et al. 2019) have shed some light on long-standing questions regarding working memory dynamics, including the rejection of burst coding. First, acros-trial voltage variability decreased through the delay, contradicting the burst coding hypothesis that predicts an increase, but supporting stable attractor dynamics (see *Neural network models of working memory, Methods*). Finally, hyperpolarizing stimulus-selective cells during the delay removed burst and spikes altogether, but kept voltage selectivity. These last results show that bursts are an intrinsic phenomena of each cell, possibly driven by spiking activity and that they are not essential to working memory.

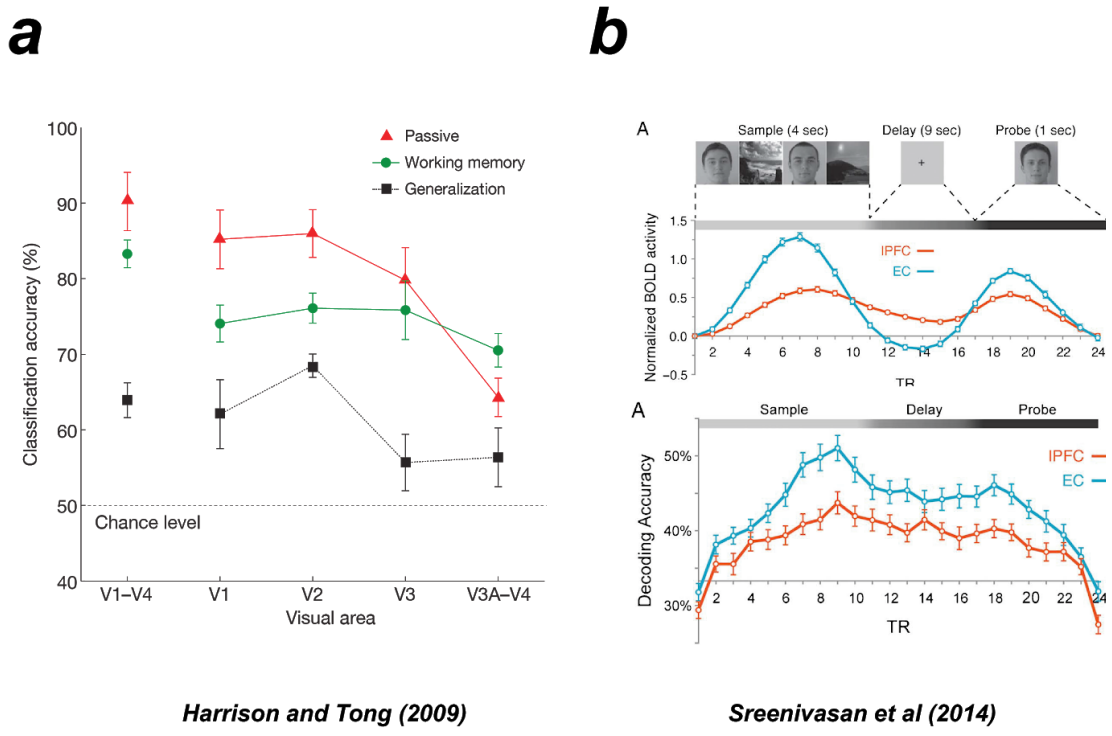
Which areas are involved: frontal vs sensory areas

After the finding of persistent activity in prefrontal cortex (PFC) (see *Persistent activity as the neural correlate of working memory, above*), some have attempted to find similar neural correlates of working memory in sensory areas. Most have failed (Figure 1.9a) or, in rare exceptions, found very weak spiking selectivity during the delay period while recording from monkeys' V4 (Chelazzi et al. 2001; van Kerkoerle et al. 2017), V1 (Supèr et al. 2001) and baboons' auditory cortex (Gottlieb et al. 1989). These null findings and the insights from early lesion studies (see Persistent activity as the neural correlate of working memory section, above) established PFC as the host area of working memory.

The advent of linear decoders applied to human neuroimaging data revealed intriguing findings that questioned the prominent role of PFC in working memory. Using linear decoders (SVM)[17], (Harrison and Tong 2009) found that it was possible to extract memory information during the delay period from areas as early as V1 (Figure 1.8a). Furthermore, they found that a decoder trained on data from subjects

[17] Support-vector machines (SVMs) are a class of supervised machine learning classifiers. Like other supervised learning classifiers, this algorithm learns the boundaries (hyperplane) between representations of two classes from a set of labeled examples.

passively viewing gratings - without any working memory - could still decode above chance when tested during the delay period of a working memory task. Delay-decoding from sensory areas was replicated by many other labs (Christophel et al. 2017). See Figure 1.9b,c for a summary of all the brain areas where delay decoding was found to be possible. Figure 1.8b shows one of these studies, where (Sreenivasan et al. 2014) decoded natural images, such as faces, from the extrastriate visual cortex (EC). Intriguingly, despite BOLD activity in PFC being higher than in EC during the delay period, multivariate decoding accuracy from PFC activity was significantly lower than from EC.



Harrison and Tong (2009) ***Sreenivasan et al (2014)***

Figure 1.8. High decoding accuracy from visual cortex BOLD signals contradicts electrophysiological findings. **a)** Decoding accuracy from several visual areas is above chance when subjects are passively looking at the stimulus (red) and while remembering it (green). Interestingly, a decoder trained on BOLD signals when the subjects were passively attending the stimuli, can extract stimulus information when they are remembering said stimuli. **b)** Despite higher normalized BOLD activity, it is easier to decode during working memory from EC than it is from lPFC.

These intriguing new findings turned cognitive neuroscientists' attention from PFC to the occipital cortex. However, BOLD signals are known to be an average of large neural populations (Dubois et al. 2015), orders of magnitude larger than the selectivity clusters found in cortical areas associated with working memory - for example, posterior parietal cortex (PPC) and prefrontal cortex (PFC) (Masse et al.

2017). Higher decoding accuracy from sensory areas - where clusters are much larger than PFC[18]- might be a reflection of this limitation in fMRI, instead of being attributable to higher information content in sensory areas relative to PFC. But one question remains: how is it possible to decode from sensory areas? As mentioned above, classical lesion studies of PFC are a key element supporting the crucial role of PFC in working memory, but we don't know what would happen if we could[19] do similar experiments with sensory areas (Scimeca et al. 2018).

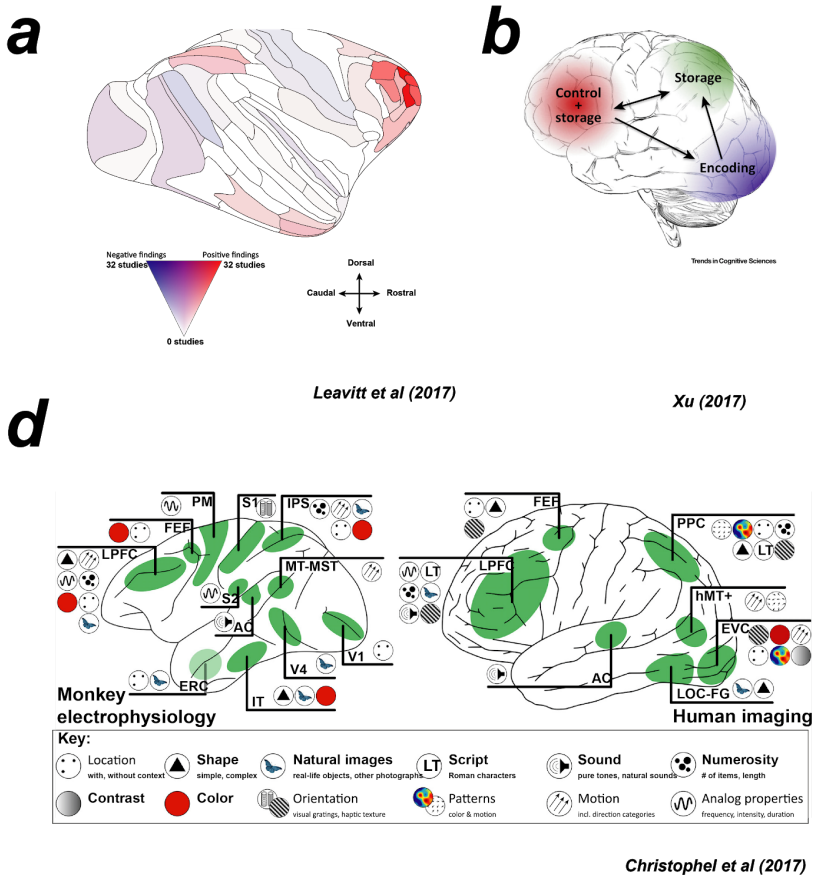


Figure 1.9. a) Meta-analyses confronting negative (blue) findings with positive findings (red) of persistent activity - reviewed also in (Christophel et al. 2017) and only for the monkey brain - reveals that robust persistent activity is actually found only in downstream areas such as posterior parietal cortex (PPC), prefrontal cortex (PFC) and inferior-temporal cortex (IT). **b)** Schematics of (Xu 2017) hypothesis. Under this hypothesis, sensory areas' involvement in working memory is constrained to stimulus encoding, meanwhile storage is confined to downstream areas. Furthermore, delay activity in sensory areas found in many studies could be explained by top-down signals coming from storage areas. **c)** Summary of all areas showing selective delay activity in monkeys (on the left) and human brain (on the right),

[18] The cat's visual cortex selectivity clusters are, for historical reasons, well characterized. See for example (Issa et al. 2000).
[19] Any working memory experiment includes perceiving the stimuli. Studies involving lesions in sensory areas are not possible because the animal would not be able to perceive the to-be-memorized stimulus.

suggesting the ubiquity of selective persistent activity. See (Christophel et al. 2017) and (Leavitt et al. 2017) for extensive reviews where these figures were taken from.

The ability to decode from the occipital cortex in the first place, some argue, could be explained by weak top-down signals coming from downstream areas such as PFC, back to sensory cortices - here called top-down hypothesis (Xu 2017). Those weak signals could be hidden in the noise of a handful of neurons recorded with electrophysiology, while whole-brain techniques such as fMRI or EEG could average the noise out of thousands of neurons, effectively capturing those weak, top-down delay signals. In fact, there is strong evidence for feedback signals in the brain. This evidence come from electrophysiological experiments (e.g. (Liebe et al. 2012; Moore and Armstrong 2003; Reinhart et al. 2012)) and neuroimaging experiments (e.g. (Bettencourt and Xu 2016; Sakai et al. 2002)) with monkeys and humans, respectively. For example, (Liebe et al. 2012) trained monkeys to perform a single-item working memory task and recorded simultaneously from V4 and PFC - both LFP[20] and single units. Crucially, during the delay period, there was enhanced communication between regions. Importantly, spiking activity in V4 was more strongly locked to prefrontal activity than vice versa, suggesting top-down communication (from PFC to V4) during the delay. Additionally, they found that the strength of inter-cortical locking - arguably a measure of communication quality - predicted the monkeys' performance, reinforcing the importance of top-down communication in working memory tasks. Moreover, there is causal evidence that the stimulation of FEF, an area bordering with PFC that is responsible for eye movements, modulates activity in V4 (Moore and Armstrong 2003). Another piece of evidence supporting the top-down hypothesis comes from decision making. Using attractor models for decision making, (Wimmer et al. 2015) showed how top-down signals from lateral intraparietal cortex (LIP) into sensory middle temporal visual area (MT) are essential to explain choice-related signals in MT, a longstanding debate in decision making (Britten et al. 1996). Although focusing on decision making, their model makes the explicit prediction that sensory areas carry working-memory information because of top-down connections. Perhaps the most direct evidence for the sensory cortex not being essential in working memory storage is the study of (Bettencourt and Xu 2016). They too could extract more stimuli information from sensory areas (V1-V4) than

[20] In contract with single-unit activity, a local field potential (LFP) is an electrophysiological signal generated by the sum of electric currents flowing through multiple nearby neurons. This signal can be obtained from low-pass filtering (~300 Hz), while single-unit activity is extracted from high-pass filtering the same signal.

from superior intraparietal sulcus (IPS) (in PPC) during the delay of a classical working memory task. However, when a distractor was introduced during the delay period, without affecting the participants' performance, it was no longer possible to decode from V1-V4, while IPS kept carrying the same amount of stimulus information. Other studies found that higher cortices filter out distractions more efficiently. See for example, (Sakai et al. 2002) for a fMRI study with humans and (Suzuki and Gottlieb 2013) for a study with monkeys comparing LIP and PFC. A recent theoretical model by (Murray et al. 2017) proposes that distractor resiliency can be accomplished by strong synaptic recurrency, which is a characteristic of higher order areas (Murray et al. 2014).

In summary, there are many reasons to believe that neural activity in sensory cortices during the delay reflects top-down influence from higher order areas such as PPC or PFC. For a recent, lively debate please refer to the series in Trends in Cognitive Sciences (Xu 2017; Gayet et al. 2018; Scimeca et al. 2018; Xu 2018).

Neural correlates of working memory capacity

There are several insights about the neural correlates of working memory load coming from human and monkey studies. Using non-invasive techniques in human subjects has revealed that BOLD and voltage amplitude - using fMRI and EEG, respectively - increased with working memory load and predicted individuals' capacity (Vogel and Machizawa 2004; Todd and Marois 2004). Moreover, both techniques have supported the slot model as signals saturate after 3-4 items (Figure 1.10a). See (Edin et al. 2009) for a theoretical model based on the bump-attractor (*Network models of working memory, Methods*) of such dynamics and (Ma et al. 2014) for alternative interpretations. Working memory load is also reflected in oscillatory power increase in high-frequency bands, such as gamma both in the monkey cortex (Kornblith et al. 2016; Honkanen et al. 2015) and in humans (Howard et al. 2003; Roux et al. 2012), while lower frequency bands such as beta (Kornblith et al. 2016; Honkanen et al. 2015), and theta/alpha (Jensen and Tesche 2002; Palva et al. 2005) seem to decrease, at least in the monkey cortex - but see (Roux et al. 2012) for an increase of alpha power with working memory load in human MEG[21]. See Figure 1.10b,c two of such studies. A recent, biophysical model of multi-item

[21] Magnetoencephalography (MEG), similar to EEG, has an excellent temporal resolution (<1 msec) but, although better than EEG, still with a poor spatial resolution.

working memory (Lundqvist et al. 2011) in which different items are stored at different phases of an ongoing oscillation, naturally explains this load-dependent power modulations (theta and gamma power increases while alpha/beta power decreases with memory load) and there is some evidence that the brain might indeed store different items at different phases, as suggested by the study of (Siegel et al. 2009). In Chapter 4.1 we developed a model that stores different items at different phases of an ongoing oscillation which power is load-modulated, as in Figure 1.10c.

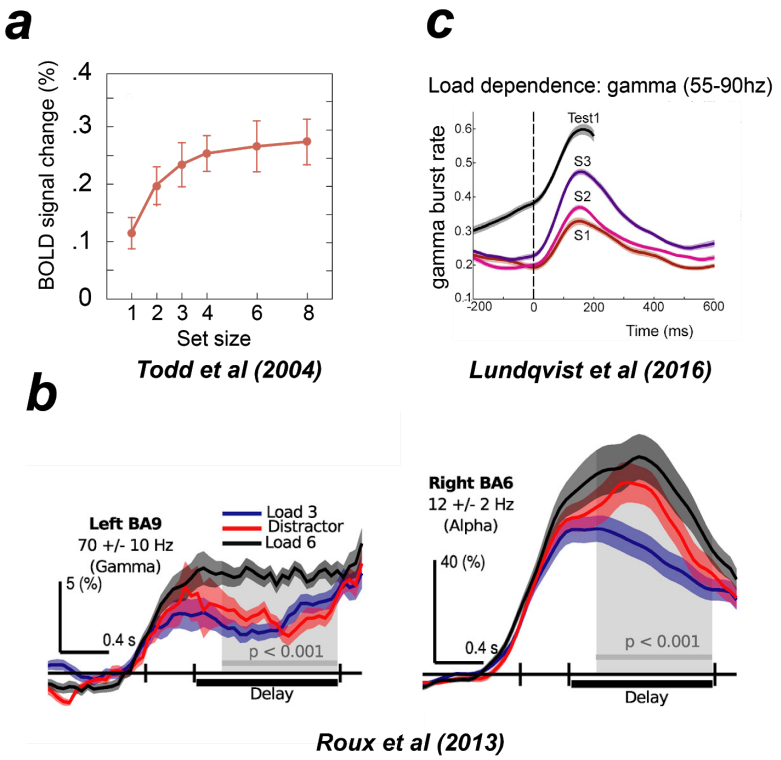


Figure 1.10. Load-dependent changes of neural activity in humans and monkeys. **a)** Human BOLD signal change during memory period relative to no memory period increases monotonically with load (set size), but might reach a plateau for set-sizes higher than 3-4 items, a signature of the slot-model (see Working memory capacity). **b)** Working memory load modulation also found in gamma and alpha power from MEG in humans and **c)** gamma from LFP in monkeys.

2. GOALS

The overall goal of this thesis is to study the neural mechanisms that underlie working memory interference, as reflected in quantitative, systematic behavioral biases. Ultimately, the goal of each chapter, even when focused exclusively on behavioral experiments, is to lay down plausible neural mechanisms that can reproduce specific behavioral and neurophysiological findings and generate predictions for future experiments. To this end, we use the bump-attractor model as our working hypothesis, with which we often contrast the synaptic working memory model. The work performed during this thesis is described here in 3 main chapters, encapsulating **5 broad goals**:

In Chapter 4.1, we aim at **(1) testing behavioral predictions of a bump-attractor network when used to store multiple items**. Moreover, we connected two of such networks aiming to **(2) model feature-binding through selectivity synchronization**.

In Chapter 4.2, we aim to **(3) clarify the mechanisms of working memory interference from previous memories**, the so-called serial biases. These biases provide an excellent opportunity to contrast activity-based and activity-silent mechanisms because both mechanisms have been proposed to be the underlying cause of those biases.

In Chapter 4.3, armed with the same techniques used to seek evidence for activity-silent mechanisms, we sought to **(4) find causal evidence for the involvement of activity-silent mechanisms in serial biases.** Finally, in light of the results from aim 4 and simple computer simulations, we **(5) reinterpret previous studies claiming evidence for activity-silent mechanisms.**

Seeking to address these aims, my thesis evolved through a constant dialogue between biophysically-constrained computational modelling (Chapter 4.1.1, Chapter 4.2.1, Chapter 4.3.2) and analyses of neurophysiological (Chapter 4.2.1) and behavioral data (Chapter 4.1.1, 4.2.1 and 4.2.2) from humans and monkeys, combined with transmagnetic stimulation (Chapter 4.3.1). This dialogue is not

finished, for our computational models provide testable predictions that motivate new, yet to be performed experiments.

3. METHODS

3.1 Neural network models of working memory

Leaky integrate and fire (LIF) neuron

There are currently many models of neuronal dynamics, each of them accounting for different levels of biological detail. For the purpose of this thesis, I focused exclusively on *leaky integrate-and-fire* (LIF) neuronal models. This model dates back to (Knight 1972) but relies on the key insight from (Lapicque 1907, later translated (Lapicque 2007)) about the relation between the neuron's membrane parameters and its excitability (Brunel and van Rossum 2007). This approach models each neuron's voltage V_i dynamics through time, $\frac{dV_i(t)}{dt}$. Because real neurons are not perfectly isolated compartments, there is a constant leak of ions (I_L) flowing through their membrane. When there is no input reaching a neuron, its voltage decays to the neuron's resting potential V_{rest} with a time constant of τ_m, following the differential equation (1). If the neuron's voltage V_i is already at V_{rest}, the change in voltage is in effect 0.

$$\frac{dV_i(t)}{dt} = -\frac{1}{\tau_m}[V_i(t) - V_{rest}] \tag{1}$$

A membrane's permeability, or *conductance* (g_L), can also be defined with respect to its capacitance (C_m) and the speed with which ions flow through its pores (τ_m): $g_L = C_m/\tau_m$

$$C_m\frac{dV_i(t)}{dt} = -I_L = -g_l[V_i(t) - V_{rest}] \tag{2}$$

Of course, any interesting neuron is embedded in a network and integrates all the inputs arriving at its synapses with other neurons. If at time t_k an excitatory neuron j spikes, the postsynaptic neuron i will increase its voltage by g_{ij}, which can be regarded as the effective synaptic strength between neuron i and neuron j. This is mathematically expressed with a function $\delta(x)$ that returns 1 when $x = 0$ (i.e. there was a spike in the presynaptic neuron) and 0 otherwise:

$$C_m \frac{dV_i(t)}{dt} = -I_{syn,e} - I_L = \sum_j g_{ij}\delta(t - t_k) - I_L$$

(3)

Crucially, when a neuron's voltage increases past a threshold V_{th}, a spike is registered and its voltage is reset to V_0. Because $V_0 < V_{rest}$, its voltage will slowly increase back to the resting potential.

Finally, we include a term to account for external stimulation of the neuron (I_{ext}) as well as for synapses with inhibitory neurons ($I_{syn,i}$), arriving at the final formalization for the LIF neuron:

$$C_m \frac{dV_i(t)}{dt} = -I_L - I_{syn,e} - I_{syn,i} + I_{ext}$$

(4)

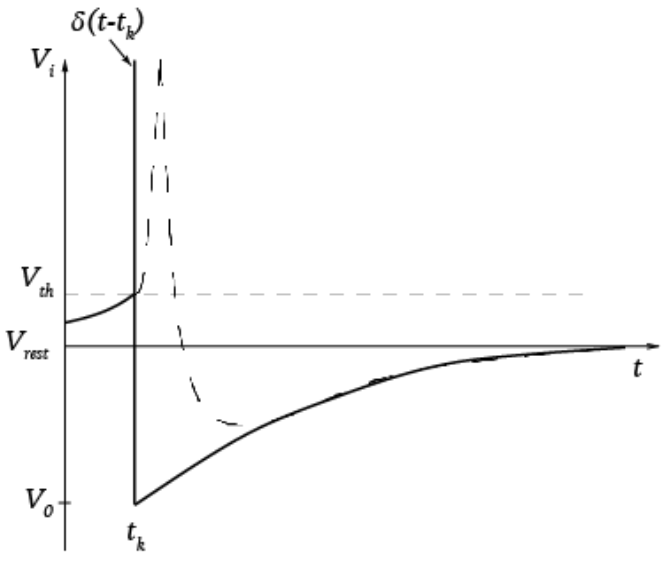


Figure 3.1. *Integrate and fire "spike".* The real shape of an action potential (dashed line) is replaced by a pulse (δ). In the integrate and fire neuron, when its voltage increases past a threshold (V_{th}), a spike is registered and its voltage is reset to $V_{0.}$ Because $V_0 < V_{rest,}$ its voltage will slowly increase back to the resting potential V_{rest}. Figure adapted from (Gerstner et al. 2014).

Conductance-based LIF

In the standard version of the LIF neuron, the effect of one neuron on its connected neighbours is instantaneous. However, it is known from electrophysiological

experiments that different channels have different opening and closing dynamics (Lester et al. 1990). For example, *AMPA* and *GABA* mediated channels are assumed to open instantaneously, but close smoothly with a time constant of τ_s. Instead of integrating its input directly through $\Sigma_k(g_{ij} \cdot s)$, with $s = \delta(t - t_k)$, we now include this smooth closing dynamics in s.

$$\frac{ds}{dt} = -\frac{1}{\tau_s}s + \sum_k \delta(t - t_k)$$

(5)

In addition, different ion channels have different permeability and let through different ions, so each of them has an associated *reversal potential*, corresponding to the membrane potential at which there is no net flow for those particular ions - i.e. the amount of ions that go in, equals the amount that go out the cell (Purves 2001). In order to account for that, we add the term $(V_i - V_r)$ - where V_r is the reverse potential for that particular type of channel. The current inflow at *AMPA* and *GABA* mediated channels of neuron i is then given by equation (6) and (7).

$$I_{i,ampa} = (V_i - V_E) \sum_j g_{ampa,ji} \cdot s_{j,ampa}$$

(6)

$$I_{i,gaba} = (V_i - V_I) \sum_i g_{gaba,ji} \cdot s_{j,gaba}$$

(7)

On the other hand, unlike AMPA, *NMDA* mediated channels have a magnesium ion at their core that blocks ion currents. When a certain level of depolarisation is reached (through other channels, such as AMPA mediated channels), this ion leaves the core and lets ions flow in. As a consequence, these dynamics slow down its opening with a time constant of τ_x. Therefore, we include this other source of dynamics through $\frac{dx}{dt}$ (equation (8)). Moreover, in addition to requiring the binding of NMDA neurotransmitters, NMDA mediated ion channels are voltage dependent. This dependence has been experimentally proven (Dingledine et al. 1999) to be well described by $\gamma(V_i) = (1 + [Mg^{2+}] \cdot exp(-0.062V_i / 3.57mV))$.

$$\frac{ds}{dt} = -\frac{1}{\tau_s}s + \alpha_s x(1-s), \qquad \frac{dx}{dt} = -\frac{1}{\tau_x}x + \sum_k \delta(t-t_k)$$

(8)

$$I_{i,nmda} = \frac{(V_i - V_E)}{\gamma(V_i)} \sum_j g_{ji,nmda} \cdot s_{j,nmda}$$

(9)

This brings us to the final differential equation describing the conductance-based LIF neuron voltage, that I will use in this thesis:

$$C_m \frac{dV_i(t)}{dt} = -I_{NMDA} - I_{AMPA} - I_{GABA} + I_{ext} - I_L$$

(10)

Short-term synaptic plasticity (STP)

The arrival of an action potential at the axon terminal of a presynaptic neuron triggers the influx of calcium ions from the cell's exterior. Consecutively, calcium ions inside the cell help containers of neurotransmitters, the *vesicles,* to fuse with the presynaptic membrane. Short-term plasticity (STP) (Stevens and Wang 1995; Markram and Tsodyks 1996; Abbott et al. 1997; Zucker and Regehr 2002; Abbott and Regehr 2004) is a biophysical mechanism in which synapses change their efficacy because of this complex presynaptic machinery. On the one hand, neurotransmitters, or neural resources, consumed during synaptic activity take time to restock, so that eventual spikes are temporarily *depressed*. On the other hand, calcium ions previously accumulated at the synapse also take time to be recycled, during which time the cell is temporarily *facilitated*. STP has been found to exhibit different dynamics, depending on cell types and cortical regions (Markram et al. 1998; Dittman et al. 2000; Wang et al. 2006).

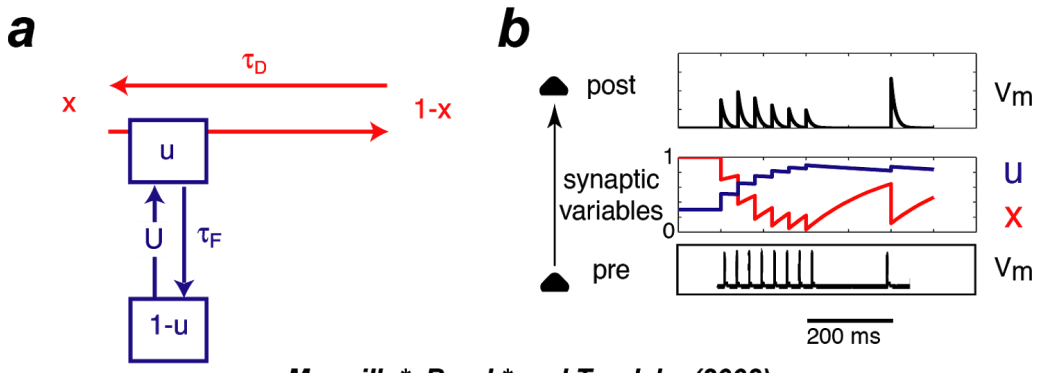


Figure 3.2. *Illustration of dynamic synapses.* ***a)*** schematic illustration of the equations 11 and 12. **b)** Example of one synapse with short-term plasticity, tuned to be a facilitated synapse. A train of presynaptic spikes has some postsynaptic responses (measured as the membrane potential, V_m) that depend on the total synaptic efficacy, which is modulated by the product ux. After each spike, u increases (facilitation) and x decreases (depression).

In a nice combination of experiments with theory, Tsodyks and Markram proposed a compact formulation that accounts for most STP dynamics (Tsodyks et al. 1998; Mongillo et al. 2008). Briefly, instead of being fixed, each synaptic efficacy has its own dynamics. Instead of being modulated by g_{ij} alone, each presynaptic spike's impact on each postsynaptic neuron (Eq. 6,7, and Eq. 11,12, Figure 3.2) is modulated by $J_{ij} = g_{ij} \cdot u_i \cdot x_i$, where u can be seen as the amount of accumulated calcium and x as the amount of available resources (e.g. vesicles) in the presynaptic neuron i.

$$\frac{du_i}{dt} = \frac{U - u_j}{\tau_F} + U[1 - u_j] \sum_k \delta(t - t_k) \tag{11}$$

$$\frac{dx_i}{dt} = \frac{1 - x_j}{\tau_D} - u_i \cdot x_i \sum_k \delta(t - t_k) \tag{12}$$

Both variables are normalized to vary between 1 and 0 in the following way. If there is no spike $(\delta = 0)$, each presynaptic variable u and x decays back to baseline values $u = U < 1$ and $x = 1$ with a time constant of τ_F and τ_D, respectively. On the other hand, if there is a spike, the amount of accumulated calcium increases by $U[1 - u_j]$, while the amount of available resources x_i, is reduced by $u_i \cdot x_i$.

Bump-attractor model

As reviewed in *Chapter 1.3, Persistent activity as the neural correlate of working memory*, many experiments have found that cortical circuits have stimulus-selective persistent activity selectivity during mnemonic periods of working memory tasks. To account for persistent activity, some sort of reverberatory connectivity is necessary. This reverberatory connectivity, or loops, could exist *between* different brain areas - for example, connectivity between PPC or thalamus and PFC - or *within*, through local recurrent connections.

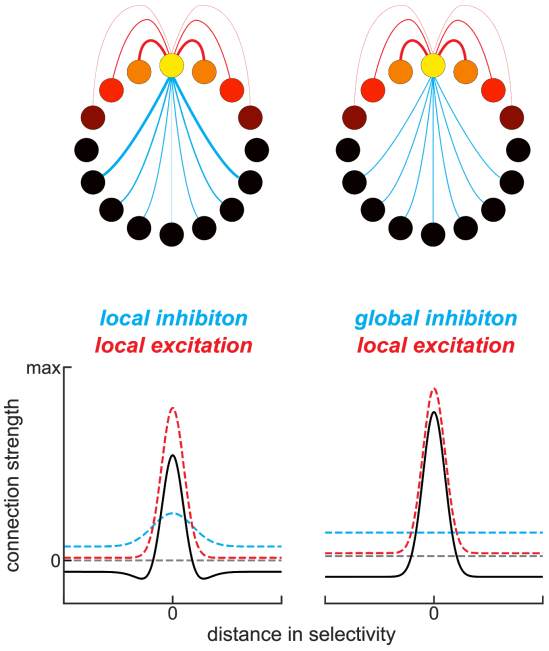


Figure 3.3. *Two types of connectivity profiles.* In this thesis, excitation will always be tuned, that is neurons with similar selectivity are more strongly connected than the ones with different selectivity. Top, each circle is a neuron and connections between them can be excitatory (red) or inhibitory (blue). The strength of each connection is represented by its thickness. On the left, local inhibition, where inhibitory neurons' connectivity is also tuned. On the right, global inhibition, where inhibitory neurons inhibit all neurons equally. Bottom, connectivity profiles of neuron marked in yellow. Black line is effectivity connectivity, which is a "mexican hat" on the left. This intermediate region of maximal inhibition predicts repulsion between memories.

In a class of attractor[22] models called bump-attractors, reverberatory activity is accomplished by translationally invariant local recurrent connection. In particular, connections between excitatory neurons (EE) coding for similar stimuli are stronger than between neurons that code for very different stimuli (Figure 3.3, red connections), which emerges as a result of prior long-term Hebbian learning (Hopfield 1982). Connections from excitatory to inhibitory neurons (EI) as well as from inhibitory to excitatory neurons (IE) can be similarly structured (Figure 3.3), but this is not necessary to have stable attractor dynamics (Compte et al. 2000; Almeida

[22] Attractor networks are dynamical networks endowed with strong recurrence. These are called "attractor" because their dynamics evolve towards, i.e. are attracted to, a stable activity pattern over time.

et al. 2015; Wei et al. 2012). On the other hand, connectivity between inhibitory neurons (II) is typically untuned - every neuron is connected similarly, despite its preferred stimulus (i.e. receptive field). Finally, broader feedback inhibition $(EI * IE)$ than recurrent excitation (EE) ensures that the network activity does not explode. Connectivity between neuron θ_i and neuron θ_j is Gaussian and defined by parameters J^+, J^- and σ (Compte et al. 2000):

$$w(\theta_i - \theta_j) = J^- + (J^+ - J^-)exp[-(\theta_i - \theta_j)^2 / 2\sigma^2]$$

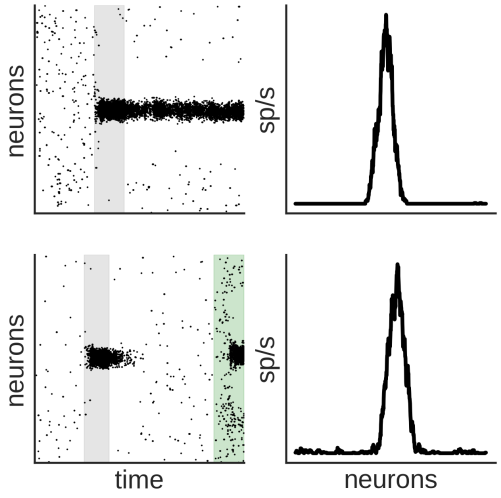


Figure 3.4. *The bump-attractor with short-term plasticity (STP) and "synaptic working memory" are essentially the same model*. On the top, a simulation of the bump-attractor with STP, on the bottom the same model, but with 10% less external input. When the input is increased at the end of the delay (green area) the bump gets reactivated (compare tuning curves at the end of the delay, on the right).

Synaptic working memory

Despite ample evidence for persistent activity, some experiments fail to find it. Moreover, recent studies (reviewed in *Silent code, dynamic code and other challenges to the stable code hypothesis, Introduction)* suggests that information that cannot be decoded from spiking activity might be recovered from a latent, silent code by means of an unspecific stimulus. One simple, yet elegant way to account for those findings is to include short-term plasticity in classical attractor networks. After the network visits one of its stimulus-specific attractors (Figure 3.4, *gray area*), short-term plasticity acting on presynaptic neurons involved in representing this attractor will bias this attractor activation, even in the absence of any attractor-specific input (Figure 3.4, green area). In other words, it is possible to temporarily store one memory in facilitated synapses, rather than in their neuron's activity. It is important to

mention, however, that having a network with embedded attractors is essential and that the bump-attractor with short-term plasticity and the “synaptic working memory” models are, essentially, the same model. Figure 3.4 shows two sets of simulations from the same model, but decreasing external current by 10% (I_{ext} in Eq. 10) for the simulations on the bottom. While simulations with an active bump drift further with increasing delays (leading to wider precision functions (Schneegans and P. M. Bays 2018), decaying synapses will eventually be reset (gradually leading to no recall), unless they are reactivated (Mongillo et al. 2008), in which case there should be no dependence on the delay duration. This prediction was tested behaviorally (Schneegans and P. M. Bays 2018), and turned out to be a validation of the bump-attractor prediction.

Nevertheless, one major advantage of including short-term plasticity, or any other subthreshold mechanism (Bliss and D’Esposito 2017; Carter and Wang 2007; Kilpatrick 2018; Hansel and Mato 2013; Mongillo et al. 2008; Barak and Tsodyks 2007), in the bump-attractor model is that this system can activate/deactivate its memories (i.e. attractors) and in-between avoid the arguable cost of spiking (Mongillo et al. 2008). Finally, it has been shown that including subthreshold mechanisms helps to stabilize memory against noise (Carter and Wang 2007).

Simulating bump reactivation

We used a previously proposed computational model (Compte et al. 2000; Edin et al. 2009; Almeida et al. 2015) to study serial dependence between two consecutive trials. The model consists of a network of interconnected 2048 excitatory and 512 inhibitory leaky integrate-and-fire neurons (Tuckell 1988). Briefly, this network was organized according to a ring structure: excitatory and inhibitory neurons were spatially distributed on a ring so that nearby neurons encoded nearby spatial locations. Connections involving excitatory presynaptic neurons were all-to-all and spatially tuned, so that nearby neurons with similar preferred directions had stronger than average connections, while distant neurons had weaker connections; while connections involving inhibitory presynaptic neurons were also all-to-all but untuned (Figure 3.3, *local inhibition*). For a detailed parameter set see *Network parameters*, below.

Short-term plasticity. Simulation of "activity-silent" mechanisms during the inter-trial period, was done by adding two more variables x and u, as described in (Mongillo et al. 2008; Markram et al. 1998) and above (Eq. 11, 12) to excitatory presynaptic neurons. The effective conductance of each excitatory synapse was then $g \cdot u \cdot x$, with g being the maximum conductance parameter.

Consecutive trials and re-ignition. Re-ignition of previous trial stimulus during the re-ignition period (300 ms before the forthcoming stimulus onset) was accomplished stimulating all excitatory neurons with a non specific external stimulus. This stimulus increased exponentially with a time constant α as $\beta(1 - e^{-\alpha(t-t_0)})$, with β being the reactivation strength. Reactivation strength was weak (0.17 *nA*) or strong (2.9 *nA*).

Network parameters

For the two-network binding model (Chapter 4.1), intrinsic parameters for both cell types were defined as in (Compte et al. 2000), as well as all the connectivity parameters, except the following:

$$G_{EE,AMPA} = 0.09\ nS,\ G_{EI,AMPA} = 0.256\ nS,$$
$$G_{EE,NMDA} = 0.24\ nS,\ G_{EI,NMDA} = 0.11\ nS,$$
$$G_{II,GABA} = 2\ nS,\ G_{IE,GABA} = 3\ nS,$$
$$g_{ext,I} = 2.74\ nS,\ g_{ext,E} = 3.5\ nS,$$
$$J_{EE}^{+} = 10,\ \sigma_{EE} = 9,\ J_{EI}^{+} = J_{IE}^{+} = 2.4,\ \sigma_{EI} = \sigma_{IE} = 18.$$

Connectivity across networks was all-to-all and untuned with the following conductances (for the unconnected simulations, these conductances were set to zero):

$$G_{EE,AMPA,across} = 0.45\ nS,\ G_{EI,AMPA,across} = 0.18\ nS,$$
$$G_{EE,NMDA,across} = G_{EI,NMDAA,across} = 0\ nS.$$

For the simulations of the bump-attractor with short-term plasticity (Chapter 4.2, 4.3) we used the short-term plasticity computational model defined in (Mongillo et al. 2008) (see *Synaptic working memory*, above), with parameters $U = 0.2,\ \tau_x = 0.2\ ms,\ \tau_x = 1500\ ms$. These dynamics affected only AMPAR mediated

recurrent connections in the network. The rest of the network parameters were as in (Compte et al. 2000) except for:

$$G_{EE,AMPA} = 0.1\ nS,\ G_{EI,AMPA} = 0.192\ nS,$$
$$G_{EE,NMDA} = 0.42\ nS,\ G_{EI,NMDA} = 0.49\ nS,$$
$$G_{II,GABBA} = 0.7413\ nS,\ G_{IE,GABBA} = 0.9163\ nS,$$
$$g_{ext,I} = 5.8\ nS,\ g_{ext,E} = 5.915\ nS,$$
$$J^{+}_{EE} = 7.1,\ \sigma_{EE} = 18,\ J^{+}_{EI} = J^{+}_{IE} = 2.2,\ \sigma_{EI} = \sigma_{IE} = 32$$

All networks included 2048 excitatory and 512 inhibitory neurons.

3.2 Behavioral Data Analysis

Behavioral datasets

For this thesis, I designed and collected behavioral data from a multi-item working memory (*Chapter 4.1.1, Dataset I*) experiment, designed and help set up a TMS experiment[23] *(Chapter 4.3.1, Dataset II),* and designed and collected behavioral data using the online-platform Amazon Turks® (*Chapter 4.3.1, Dataset IV*). Additionally, I also downloaded and put together a large dataset consisting of 8 open datasets (*Chapter 4.2.2, Dataset III*).

Dataset I: Multi-item working memory

The experimental paradigm is schematically illustrated in Figure 4.1.2. Stimuli presentation was followed by a delay period of 3 s. After the delay period, the fixation dot changed from black to the color of one of the previously presented items. The subject was required to respond by indicating the remembered position of the item matching the color of the fixation mark. To indicate the remembered position, the subjects used a pressure-sensitive tablet and pen. The movement of the pen was reproduced in the visual display as a cursor so that the subjects saw the colored fixation dot moving from the fixation spot to the remembered position. The subject indicated the reported position by releasing the pen from the tablet. All trials had a delay of 3 s and separation between nearby items ranged from 3.1 to 4.4 deg of

[23] Data collection and performed entirely by Rebecca Martinez, who was also crucial for setting up the experiment and analysing and discussing the data.

visual angle (14 to 20 deg on the circle). Data was acquired from 4-8 sessions from each of 9 healthy participating subjects (4 females), ages between 21 and 27 years old and showing normal or corrected-to-normal vision. For each subject, sessions were typically acquired on different days. Some participants completed fewer sessions, because they were not available for more data collection. The trials where the probed item was near another item were classified into two trial types, according to the probed item being *clockwise* or *counter-clockwise* relative to the nearby item.

Dataset II: TMS experiment dataset

Human participants and behavioral tasks. Twenty (20) neurologically and psychologically healthy volunteers with normal or corrected vision (n=20, 29.86 years ± 9.55 years (mean ± std)) from the Barcelona area provided written informed consent and were monetarily compensated for their participation, as reviewed and approved by the Research Ethics Committee of Hospital Clínic (Barcelona). Each participant performed two sessions of approximately 1.5 h within 24 h. Stimuli were presented on a 17'' HP ProBook using Psychopy (version 1.82.01) at a distance of 65 cm from the participant's eyes. This study consists of an original experiment with 10 subjects, and a preregistered replication experiment (https://osf.io/rguzn/) with 10 more subjects.

Each trial began with the presentation of a central black fixation dot (.5 x .5 cm) on a grey background. After 1 s of fixation, a single black circle (stimulus, diameter 1.4 cm) appeared for .25 s at any of 360 circular locations at a fixed radius of 4.5 cm. The stimulus was followed by a 1 s delay in which only the fixation dot remained visible. A change of the fixation dot color from black to red instructed participants to respond (response probe). Participants responded by making a mouse click at the remembered location. A transparent circle with a white border indicated the stimulus' radial distance, so the participant was only asked to respond with its angular location. After the response was given, the cursor had to be moved back to the fixation dot to start a new trial. Participants were instructed to maintain fixation during pre-stimulus fixation, stimulus presentation, and delay and were free to move their eyes during response and when returning the cursor to the fixation dot. Angular positions (1 out of 360), were randomly sampled from a uniform distribution at the beginning of each session. At the end of the fixation period, a single pulse of TMS was applied in half of vertex trials (weak/strong tms or no-tms), and in two thirds of prefrontal trials (weak,

strong or no-tms). Participants completed 4 blocks of 90 (vertex) or 130 (PFC) trials within each session.

Transcranial Magnetic Stimulation. Stimulation was performed in the TMS study using a Magstim Rapid 2 machine with a 70 mm figure-of-eight coil. TMS target points were located using a BrainSight navigated brain stimulation system that allowed coordination of the coil position based on the participant's structural MRI (sMRI) scan. A region of interest in dlPFC was defined using NeuroSynth term-based meta-analysis of 53 fMRI studies associated with the key phrase 'spatial working memory'. This mask was transformed into each subject's sMRI space. Vertex target points were defined using a 10-20 measurement system. Stimulator intensity, coil position, and coil orientation were held constant for each participant for the duration of each session. In order to mask the sound of TMS coil discharge, we had participants listen to white noise for the duration of the session. White noise volume was selected based on participant threshold for detecting TMS click using the staircase method (2-up, 1-down). Stimulation intensity was determined by the individually-defined resting motor threshold (RMT). The stimulation parameters were in accordance with published TMS guidelines (Rossi et al. 2009). .

Dataset III: Open datasets

We analysed 8 datasets that are freely available online (*Chapter 4.2.2*, Table 3.1), with a total of n=760 subjects performing variations of the same, delayed estimation of color task (*Chapter 4.2.2*, Figure 4.2.9a). We will briefly describe the general experiment and Table 3.1 summarizes the specifics of each task, for detailed descriptions please refer to the original studies (Foster et al. 2017; Souza et al. 2014; Oberauer and Lin 2017; Bays et al. 2009; van den Berg et al. 2012). On each trial, a set of colored stimuli (varying from 1 to 8 stimuli) were briefly shown. After a delay period of roughly 1 second (see Table 3.1 for details), during which stimuli were no longer visible, subjects had to report the target color of a cued location. These color reports correspond to angles (i.e. degrees) on a color wheel rotated by a random amount on every trial, to avoid a spatial memory strategy.

Dataset	*Set size*	*Subjects*	*Trials*	*Delay*	*Observations*
CamCan data set Taylor et al (2017) Shafto et al (2014)	1-4	649	224	0.9 s	Stimuli: circle of diameter 1.77 (dva), positions selected at random from 8 equally spaced points at an eccentricity of 4.5. 360 colors. CIE L,a,b radius of 53 and center (64,10,10). Half of the trials had non-targets probing features revealed. These data was obtained from the CamCAN repository, available at: www.mrc-cbu.cam.ac.uk/datasets/camcan/
Experiment 1 of Oberauer & Li (2017)	1-8	19	400 x 2	1 s	Stimuli: colored squares of 1.25º at viewing distance of 50cm. 360 different colors on a color wheel: CIE L,a,b = (70,20,30).
Experiment 1 (I) of Foster et al (2017)	1	12	~960	1.2 s	Stimuli: circle, 1.6º diameter, centered at 3.8º at viewing distance of 100cm. 360 colors. Color wheel in Figure 4.2.9
Experiment 2a (II) of Foster et al (2017)	1	21	~960	1.15 s	Stimuli as I, plus: During presentation, a distractor with different shape and color was present at another location.
Experiment 1 (I) of Van den Berg et al (2012)	1-8	13	288 x 3	1 s	Stimuli: circle, 2º diameter, centered at 4.5º at viewing distance of 60cm. 180 different colors on a color wheel: CIE 1979, L,a,b = (70,10,10).
Experiment 3 (II) of Van den Berg et al (2012)	1-8	13	288 x 3	1 s	Same as I, but report done by scrolling through all possible colors (drawn uniformly and independently from the wheel).
Experiment 2 of Souza et al (2014)	1-8	21	496 x 2	1 s	Stimuli: circles of 1.1 cm diameter at 5.5cm from fixation. 360 different colors samples from hue dimension of HSL (saturation=1, lightness=.5). Condition 1: color wheel present during delay Condition 2: Last 1 sec of - sec delay was location cued.
Bays et al (2009)	1-8 (in blocks)	12	600	0.9 s	Stimuli: 2x2º patches at viewing distance of 57cm. 180 different colors samples from CIE L,a,b=(70,20,38). Presentation duration varied: 0.1, 0.5, or 2 s in a block design. This could potentially confound build-up analysis.

Table 3.1. *Experimental details of each dataset.* With the exception of Foster et al I & II (Foster et al. 2017), all datasets have a varying set size.

Dataset IV: Amazon-Turks dataset

We designed an experiment to be run in the online platform Amazon Mechanical Turks in order to obtain data from a large sample of participants and thus increase the statistics for what we expected to be a small effect (Chapter 4.3). The experiment was similar to Dataset II, but instead of using TMS, we employed visual stimulation by changing the background color between white and gray at two different frequencies (5 hz or 10 hz) during the last 500 ms of the pre-cue period. A total of n=237 started our experiments, but only n=112 passed our screening filter, which consisted of selecting only subjects which session had at least 100 correct trials.

Mixture-model fitting and statistical model comparison

To test alternative statistical models the data was fitted to three statistical models detailed below using a custom expectation maximization algorithm for the maximum likelihood estimation based on publicly available code ((Bays et al. 2009) http://www.paulbays.com). Model comparison was done using Akaike information criterion (AIC) (Burnham and Anderson 2004), which is a measure of the relative quality of a statistical model for a given data set. Information loss of one model relative to another is then calculated by the differences between AIC values (Burnham and Anderson 2004). The information loss ΔAIC_i of each model compared to the best (the one with the lowest AIC) was calculated for each subject and then averaged across subjects. The relative likelihood of model i relative to the best model was computed as $exp(\Delta AIC_i / 2)$.

Swap model. This model is the one introduced by (Bays et al. 2009), to account for performance on a recall task where both stimuli and responses are chosen from a circular parameter space. The model assumes that the experimental distribution can be described as a mixture of 3 components:

$$f_{exp}(\Delta\theta) = p_t\phi_\sigma(\Delta\theta) + p_{nt}\frac{1}{n}\sum_i \phi_\sigma(\Delta\theta_i^*) + p_u\frac{1}{2\pi} \quad (13)$$

Attraction model. In this model the subjects' reports are described by a unimodal von Mises distribution centered on a location intermediate between the target and non-target items. This displacement would occur as a result of the attraction of coding bumps in our more detailed model of Figure 4.1.1. This model drops one of the typical components (Bays et al. 2009), the possibility of having swap errors, and introduces a bias b in the mean, representing the attraction effect:

$$f_{exp}(\Delta\theta) = p_t\phi_\sigma(\Delta\theta + b) + p_u\frac{1}{2\pi} \tag{14}$$

Attraction+swap model. Finally, both swaps and attraction biases might co-exist: in some trials the two features of the stimulus are misbound, but in any case reports (to target or to non-target items) are biased towards the nearby stimulus. This model is the same as the *swap model* but with one more parameter for the bias:

$$f_{exp}(\Delta\theta) = p_t\phi_\sigma(\Delta\theta + b) + p_{nt}\frac{1}{n}\sum_i \phi_\sigma(\Delta\theta_i^* - b) + p_u\frac{1}{2\pi} \tag{15}$$

Serial biases

Human. For each trial, we measured the response error (θ_e) as the angular distance between the angle of the presented stimulus and the angle of the response. To exclude responses produced by guessing or motor imprecision, we only analyzed responses within an angular distance of 1 radian from the stimulus.

We measured serial biases as the average error in the current trial as a function of the circular distance between the previous and the current trial's target location (θ_d, *prev-curr*) in sliding windows with size $\pi/3$ and in steps of $\pi/20$. To increase power and correct for global response biases, we calculated a 'folded' version of serial biases as follows. Trial-wise errors were multiplied by the sign of *prev-curr* distances: $\theta_e = \theta_e * sign(\theta_d)$ and we only used absolute values of *prev-curr.* Positive mean folded errors should be interpreted as attraction towards the previous stimulus and negative mean folded errors as repulsion away from the previous location. For absolute serial bias analyses we average folded errors up to π.

Monkey. In contrast with the human experiments, the stimulus distribution was discrete for all the monkey experiments. On each trial, the subject was cued for 1 out of 8 possible cue locations equidistant on a circle. This restricted the minimum distance between two consecutive trials to be $\pi/2$. To have a finer resolution to calculate serial biases, we capitalize on the response variability on each trial: we computed relative distances between current trial stimulus and previous trial response (instead of previous trial stimuli). Similar methods to humans were used, where we used smaller sliding window sizes ($\pi/10$ at steps of $\pi/100$), essential to capture the thinner attractive profile we saw in monkeys.

Regression models

To test the effect of TMS on serial biases, we fit a linear mixed-effects model using R function *lmer* (Bates et al. 2015). In particular, we modeled trial-wise errors as a linear combination of the coil *location (PFC vs. vertex)*, tms *intensity (strong-tms, sham, and weak-tms)* and the sine of the distance between the previous and current stimulus (*prevcurr)*. We accounted for subject-by-subject variability by including random intercepts and random coefficients of *prevcurr*. Moreover, we incorporated the non-linear dependency of serial bias on stimulation intensity that surfaced in our simulations, by using -1, 0 and 1 for strong-tms, no-tms and weak-tms, respectively. The full, three-way interaction model was:

$$\theta_e \sim location * intensity * prevcurr + (1 + prevcurr \mid subject) \tag{14}$$

3.3 Neural Data Analysis

Datasets

For this thesis, I analyzed one neural dataset previously collected by (Constantinidis et al. 2001b), briefly described below. Additionally, I also analysed simulated neural data from variations of the bump-attractor model. Description of such analyses are at the end of this chapter.

Neural dataset and accompanying behavioral task

Detailed methods of the behavioral task, training, surgeries and recordings, as well as descriptions of neuronal responses in the task have been published previously (Wimmer et al. 2014; Constantinidis et al. 2001a; Compte et al. 2003; Constantinidis et al. 2002; Constantinidis and Goldman-Rakic 2002) and are only summarized briefly here. Four adult, male rhesus monkeys (Macaca mulatta) were trained in an oculomotor delayed response task requiring them to view a visual stimulus on a screen and make an eye movement after a delay period. During execution of the task, neurophysiological recordings were obtained from the lateral prefrontal cortex. Visual stimuli were 1° squares, flashed for 500 ms at an eccentricity of 14° relative to the fixation point. Stimuli were presented randomly at one of 8 possible locations around the fixation point. A delay period lasting 3 s followed the presentation of the stimulus, at the end of which the fixation point turned off, and an eye movement towards the location of the remembered stimulus was reinforced with liquid reward. Eye position was monitored with a scleral eye coil system. From two of those monkeys, we collected single-unit responses from dorsolateral PFC (dlPFC) using tungsten electrodes of 1–4-MΩ impedance at 1 kHz, while they were performing the task. A substantial fraction of neurons in this area showed tuned persistent delay activity during the mnemonic delay period of the task (n=206/822, (Wimmer et al. 2014; Constantinidis et al. 2001a; Compte et al. 2003; Constantinidis et al. 2002; Constantinidis and Goldman-Rakic 2002)). For decoding analyses, we grouped those neurons in simultaneously recorded ensembles (total of n=94 neural ensembles). All experiments were conducted in accordance with the guidelines set forth by the US National Institutes of Health, as reviewed and approved by the Yale University Institutional Animal Care and Use Committee.

Preferred location

From neuronal spike counts, we computed the preferred locations of each neuron. Similar to (Wimmer et al. 2014), preferred location was determined by computing the circular mean of the cue angles (0° to 315°, in steps of 45°) weighted by the neuron's mean spike count over the delay period (3 s) following each cue presentation.

Cross-correlation

Dataset. For the estimation of functional connectivity we computed jittered cross-correlation (Amarasingham et al. 2012) of spike counts from simultaneously recorded neuron pairs, whose preferred locations were separated by a maximum of 60º (n=74). For each pair we selected the trials that fell within the preferred range (pref, max of +40º from each preferred locations) or outside the preferred range (anti-pref, all other trials). We discarded trials without at least 1 spike per second.

Jittered cross-correlation. We used the Python function scipy.signal.correlate to compute cross-correlations between spike trains of simultaneously recorded pairs. Spikes were counted in independent windows of 10 ms. From each trial cross-correlation, 1000 jittered versions were computed as follows (Amarasingham et al. 2012). We shuffled the spike counts on windows of 50ms and computed cross-correlation for each of these jittered versions. This captured all the cross-correlations caused by slow dynamics (>50ms) but destroyed any faster dynamics. Finally, we removed the mean of 1000 jittered versions of each trial, leaving only correlations caused by fast dynamics (<=50ms). Before plotting cross-correlation peaks, we averaged 3 bins (+1/-1 bin from the actual peak). For the time resolved cross-correlation, we repeated this process for sliding windows of 1 s and steps of 50 ms.

Putative excitatory and inhibitory interaction. Based on the average cross-correlation peak during the whole trial [-4.5,2.5], we classified each pair into one of 3 subgroups: 1) those with a positive peak for both preferred and anti-preferred trials were classified as excitatory, 2) those with negative peak for both preferred and anti-preferred trials were classified as inhibitory and 3) we discarded those with inconsistent peak sign between preferred and anti-preferred trials. In total, we analyzed the cross-correlation time course of n=47 pairs of neurons (n=27 excitatory and n=20 inhibitory).

Decoding stimulus information

Monkeys.

Population decoder. We decoded stimulus θ_j in trial j by modelling it as a linear combination of the spike counts of simultaneously recorded neurons n_i, computed in sliding windows of 0.5 s and steps of 0.1 s during that trial.

$$cos(\theta_j) \sim 1 + \sum_{i}^{k} \beta_i n_i \quad , \quad sin(\theta_j) \sim 1 + \sum_{i}^{k} \omega_i n_i \tag{15}$$

We performed 50-fold cross-validation. For each set of neurons, we trained two sets of weights β_i and ω_i on 80% of randomly selected trials and tested in the remaining trials.

Leave-one-out decoder. To measure stimulus information on a trial-by-trial basis, we used leave-one-out cross validation. We regressed the β_i and ω_i weights in all trials, except the one left out for testing. For this analysis we computed spike counts in windows of 1 s in steps of 50 ms.

Distance from shuffle. For our final measure of decoding accuracy, z, we compared each ensemble decoding accuracy, acc, to that ensemble's decoding accuracy in 1000 shuffled stimulus distributions, acc_{shuff}. By shuffling the distribution of stimuli presented in the particular recording of each ensemble, we maintained the characteristics of the distribution (e.g. unbalanced distribution of stimuli), but effectively destroyed correlations between stimuli and neural activity.

$$z = \frac{acc - mean(acc_{shuff})}{std(acc_{shuff})} \tag{16}$$

Humans[24]*.*

Linear decoder. EEG alpha power is known to decrease in occipital sites contralateral to attended locations and locations being actively maintained in working memory (Worden et al. 2000; Kelly et al. 2006; Medendorp et al. 2007; Foster et al. 2016). We used this feature to decode the stimulus' angular position from the distribution of alpha power over all 43 electrodes. As for the monkey data, we used a linear leave-one-trial-out decoder trained on the previous trial's stimulus label and decoded this information throughout the previous and current trial. Trialwise alpha power for each electrode was modeled as a linear combination of a set of regressors representing the stimulus location in the corresponding trial, $U = WM$, where U is a $J \times K$ matrix of alpha power measured at electrode j in trial k, M is the $N \times K$ design matrix of values for regressor n in trial k, and W is the $J \times N$ weight matrix, mapping the weight for regressor n to electrode j.

Design matrix M. The design matrix M is a set of eight regressors M_n representing expected "feature activations" (Brouwer and Heeger 2009) for feature n in trial k. The value of regressor M_n in trial k was determined as $|sin(n\pi/8 - s_k\pi/8 + \pi/2)^7|$, where $s_k = [0 .. 7]$ indicates one of eight angular location bins corresponding to the stimulus shown in trial k.

Similar to monkey analyses, we measured single-trial stimulus representations using leave-one-out cross-validation, using an equal number of trials from each location bin in the training set (U_{train} and M_{train}). We estimated the weight matrix $\hat{W}$ by $\hat{W} = U_{train}{M_{train}}^T(M_{train}{M_{train}}^T)^{-1}$ and estimated the left-out trial k's design matrix M_k as $\hat{M}_k = (\hat{W}^T\hat{W})^{-1}\hat{W}^T U_k$.

For each trial and time point, we ran repeated this analysis 100 times with randomly chosen training sets, and averaged $\hat{M}$ over all repetitions. Finally, we estimated the predicted angle $\hat{\theta}_k$ as the direction of the sum of vectors $\vec{v}_n$ with length $\hat{M}_{nk}$ and

[24] All the EEG data analyses were performed entirely by Heike Stein, a fellow PhD student in the lab. Because I was involved in discussing the results during and after the EEG analyses, and because those results complement substantially our findings in the monkey PFC, I opted to include some of them here for completeness' sake. This methods section is included here to properly interpret those findings.

direction b_n. Trialwise decoding strength was then given as the cosine of the circular distance between $\hat{\theta}_k$ and the actual stimulus angle in that trial θ_k.

Cross-temporal decoding. To explore the temporal generalization of mnemonic and response code over time, we trained decoders in independent time windows of the previous and current trial, and tested them in all time points of consecutive trials (from .25 s to 1.25 s after previous stimulus onset, -.25 s to .25 s after previous response, and -1.25 s to .25 s after the current trial's stimulus onset). For the temporal generalization matrix (Figure 4.2.4b), we averaged train and test data over 50 samples (≈ 97.77 ms). High-resolution time courses of mnemonic and response code (Figure 4.2.4c) were obtained by training the decoder on averaging data from .5 s to 1 s after previous stimulus onset and -.25 s to .25 s relative to response time, and testing on averaged data from five samples (≈ 9.77 ms) throughout consecutive trials.

Conversion of spikes into local-field potentials

For the conversion of simulated spike times into local-field potentials, we convolved the aggregated spike times (t_s) of all the neurons engaged in a bump (or in the network, depending of the analyses) with a synaptic kernel g_{syn} :

$$g_{syn}(t) = \overline{g}_{syn} exp(-\frac{t - t_s}{\tau}) \tag{17}$$

that had an exponential decay of $\tau = 5\ ms$, as defined in (Sterratt et al. 2011).

Phase-preservation index

To measure how an oscillating bump kept its oscillation phase over multiple trials *N of* our simulation, we first converted spike times into local-field potentials, and then we used the phase-preservation index (PPI), a method developed by (Mazaheri and Jensen 2006) for MEG data. The PPI is defined by taking a reference time point (in our case t_{ref} = stimulus offset), and then computing the average consistency of the phases at a specific frequency of interest f_0 across the rest of time points:

$$PPI(f_0,t) = \frac{1}{N}\sum_{k=1}^{N} |e^{i\phi^k(f_0,t_{ref})-i\phi^k(f_0,t)}|$$

(18)

where $\phi^k(f_0,t_{ref})$ is the reference phase for trial k and $\phi^k(f_0,t)$ is the phase in trial k at all times t in the frequency of interest f_0.

4. RESULTS

4.1 Interference from simultaneous memories

Neural circuit basis of visuo-spatial working memory precision[25]

Summary

Here we used a neuronal microcircuit model for visuospatial WM (vsWM) to investigate working memory of several items. The model assumes that there is a topographic organization of the circuit responsible for spatial memory retention. This assumption leads to specific predictions, which we tested in behavioral experiments. According to the model, nearby locations should be recalled with a bias, as if the two memory traces showed attraction or repulsion during the mnemonic period, depending on their distance. We confirmed these predictions experimentally. Our findings provide support for a topographic neural circuit organization of vsWM, they suggest that interference between similar memories underlies some WM limitations, and they put forward a circuit-based explanation that reconciles previous conflicting results on the dependence of WM precision with load.

[25] This chapter includes parts of a study published in 2015: Rita Almeida, João Barbosa, and Albert Compte. “Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study” Journal of Neurophysiology 114:1806, 2015. From that study, I included here only the analyses and experiments in which I was directly involved.

In computational models, simultaneous memories attract and repel each other

Continuous-attractor networks can store several items in separate activity bumps (Edin et al. 2009). However, the localized activity bumps that represent these memories interfere, reflecting the connectivity profile of these networks. Figure 3.3 (Methods) illustrates two of such profiles, local and global inhibitory connectivity. Both connectivity profiles predict that similar, simultaneous memories are attracted to each other during the mnemonic period (Wei et al. 2012; Almeida et al. 2015). Local inhibition, however, predicts that dissimilar memories repel each other in the course of one trial (Figure 4.1.1a, (Almeida et al. 2015; Nassar et al. 2018)). When two items are stored in a network with local inhibition, there is a range of distances where the two bumps merge. This range depends on the width of the excitatory-to-excitatory connectivity profile (Figure 3.3). On the other hand, repulsion depends on the width of the total feedback inhibition. As shown in Figure 4.1.1b, the interaction between two nearby memories transitioned from attraction to repulsion as the inter-item distance grew.

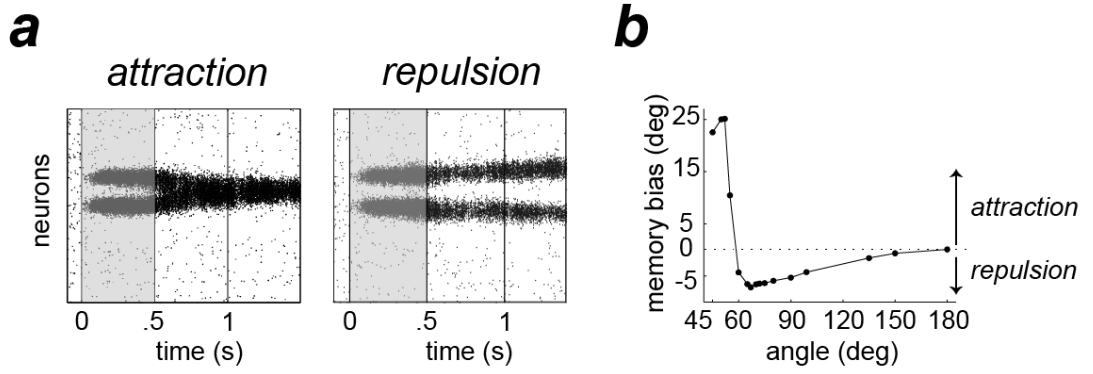


Figure 4.1.1. *Simultaneous memories interfere attractively and repulsively.* **a)** Two example simulations with slightly different iter-item distance. On the left, bumps are placed nearby and therefore merge. On the right, bumps are placed slightly further away from each other, and therefore repel. For each trial we computed the memory bias, as the distance between the initial bump location during stimulus and the location at the end of the delay (last 0.5 s). In **b)** we plotted memory biases for all inter-item distances and the folded Mexican hat imposed in the network connectivity emerges.

Testing bump-attractor model predictions in humans

We designed a multi-item working memory experiment (Figure 4.1.2, see *Dataset I* in *Methods* (3.2) for more details) with a parametric report to test the attraction and

repulsion predictions (Figure 4.1.1). In this task, nine participants had to report the remembered locations by controlling a cursor. We focused our analysis on *close trials*, which were characterized by reports to a target stimulus that had been memorized together with a non-target stimulus presented nearby, within 3.1 to 4.4 degrees of visual angle (14 to 20 deg on the circle). Depending on whether this close non-target was presented clockwise or counter-clockwise to the target, we named each close trial a *cw* or *ccw* trial, respectively. We found that there was a significant difference between the reported errors for the *ccw* and *cw* trial types (Figure 4.1.3a, $p < 0.0001$). This data was consistent with attraction of the two memories. We were able to measure the specific fraction of a perfect merge verified in the data. We did this by normalizing the mean error in each trial to the distance between close stimuli. The subjects that showed a significant effect (5 out of 9) presented 26% ± 8% (39% ± 6%) of the attraction expected for a total merge of the memories in clockwise (counter-clockwise) trials.

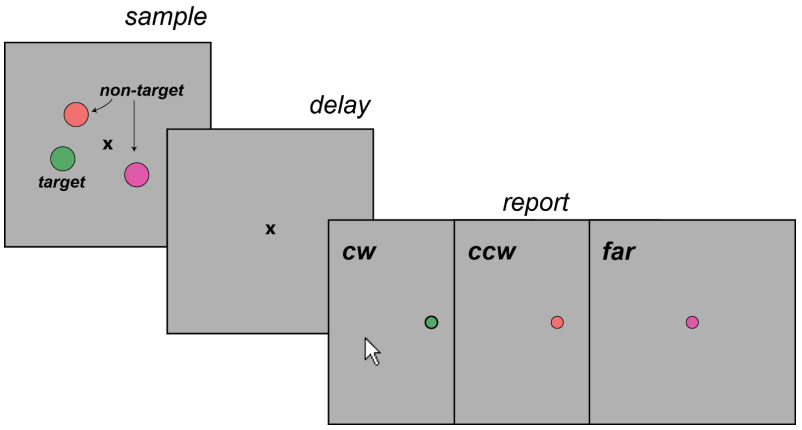


Figure 4.1.2. *Task design used to test the bump-attractor predictions in humans.* In each trial, the subject saw 3 items placed at different locations each with distinguishable colors. On each trial, there was 1 target and 2 non-targets labels invisible to the subjects (shown only for illustration). Because subjects were unaware of this target/non-target categorization, they had to remember all items during a delay period of 3 seconds, at the end of which the target location was probed by the color change of the fixation dot. Finally, the subjects used a mouse to report the target location, which could have a non-target next to it (close trial, cw or ccw trial) or not (far trial).

Moreover, we computed the memory bias from the psychometric curve fit for each subject and plotted it against distance between items (Figure 4.1.3c). Across subjects, the attractive memory bias of the psychometric curve decreased significantly (1-tailed paired t-test, $p = 0.02$; $n = 9$) from very close memories (3.0 –3.5° of visual angle, memory bias 95% confidence interval [0 0.7]°, permutation test

p = 0.05) to slightly more distant ones (4.2° of visual angle), at which point the memory bias became marginally negative (memory bias 95% confidence interval [-1.2 0.1]°, permutation test p = 0.07). In addition, we tested significant memory biases within subjects. We found that the number of subjects with a significant repulsive (attractive) memory bias increased (decreased) with distance between items (Figure 4.1.3d; multinomial regression model p = 0.035), indicating a consistent but individually specific dominance of repulsion for intermediate distances.

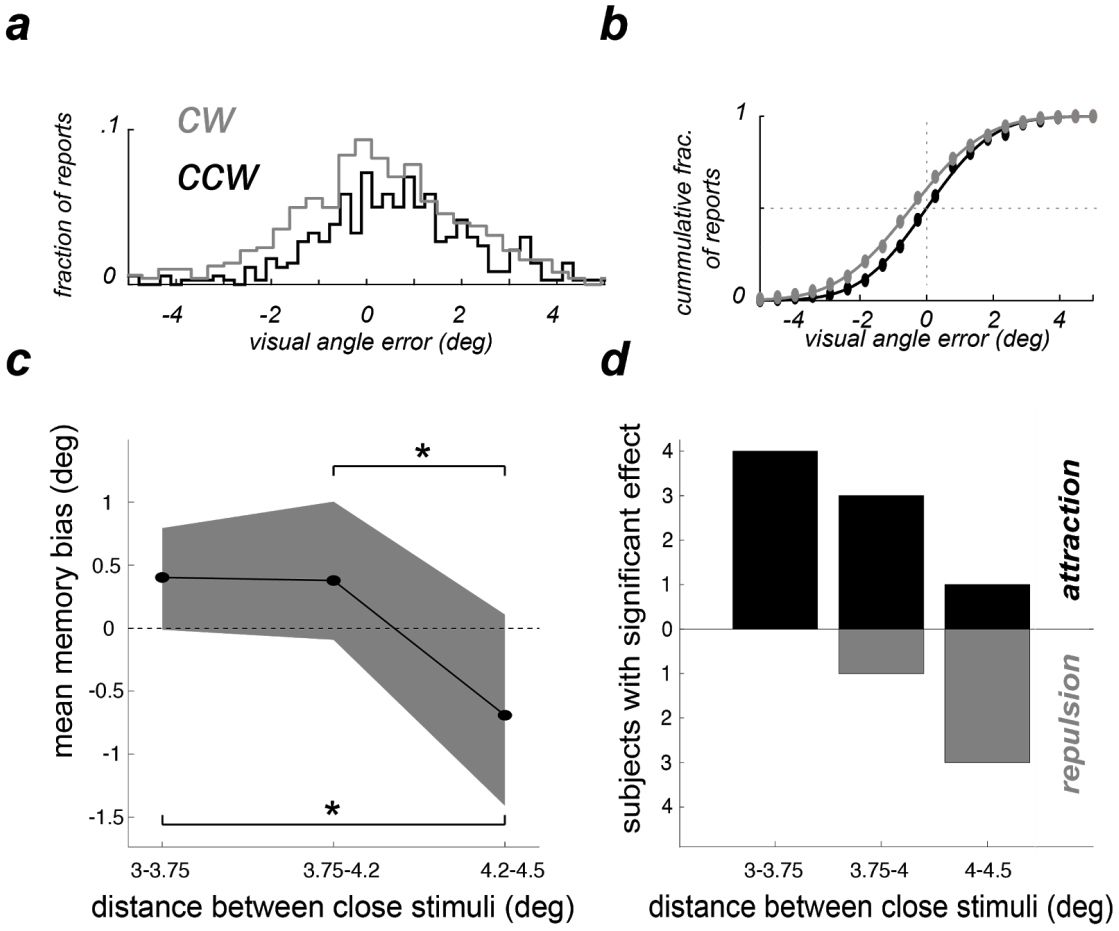


Figure 4.1.3. *Memory attraction and repulsion emerges depending on distances between close-by items.* **a)** distributions of error to target for clockwise (gray) and counterclockwise (black) trials differed significantly (p = 0.00005, data from all participants n = 9), revealing an attractive bias. **b)** Cumulative proportion of errors to target from the distributions. Data were fitted with a cumulative normal function**. c)** subject-averaged memory bias for trials with different distances between memorized close-by items (x-axis). Shading indicates bootstrap-derived 95% confidence intervals. *Significant difference as evaluated with 1-tailed paired t-test at p < 0.05. **d)** no. of subjects with significant (t-test p < 0.05) attractive and repulsive memory bias in trials with different interitem distance.

Discarding swap-errors as an alternative explanation

An alternative explanation to the attraction between similar items could be that, in a fraction of error trials, the subjects swapped the colors and locations of the two memorized nearby items (Bays et al. 2009; Ma et al. 2014; Pertzov et al. 2012), while for the rest of error trials subjects were randomly guessing (see *Binding of independent features* in *Introduction (1.2))*. Misremembering the binding between color and location would result in a spurious attraction - in fact, it would be a complete attraction.

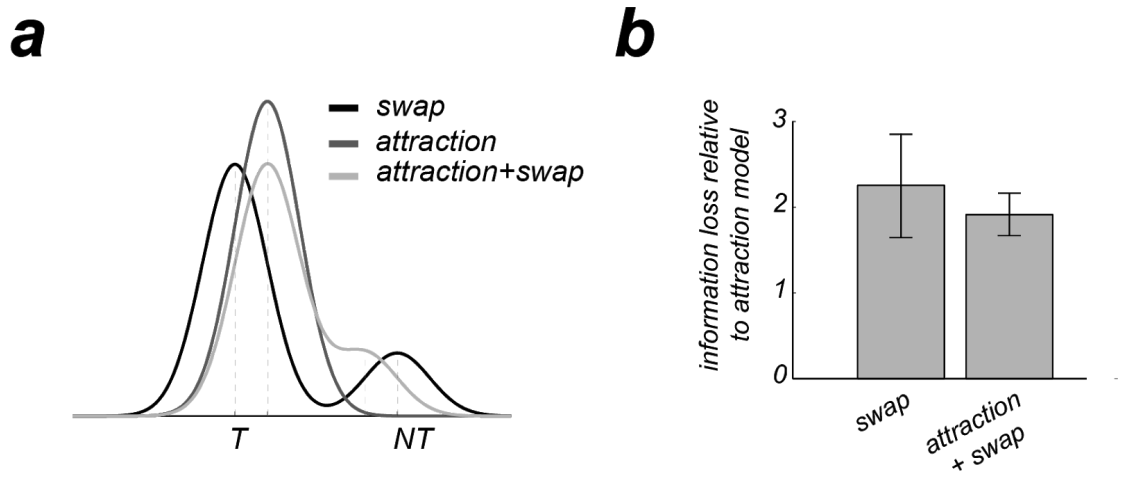


Figure 4.1.4. *Behavioral data suggest that attraction of memory representations and not swap errors is responsible for memory biases observed in close trials.* **a)** Schematic illustration of the probability density function for each of the 3 models tested: swap, attraction, and attraction+swap models. **b)** average information loss ΔAIC across subjects (n = 8) for swap and attraction+swap models compared with the attraction model, the best model for data from these participants.

We fitted behavioral reports with statistical models that included Gaussian-like distributions around the target memory items (*Mixture-model fitting* in *Methods)* using a custom expectation maximization algorithm based on (Bays et al. 2009). For all tested models, the dispersion parameter σ estimated from trials with close probed items (σ = 7.63 ± 0.88 deg along the circle, *n*=9) did not differ significantly from that estimated from trials with far probed items (paired t-test, *p*> 0.05, *n*=9), suggesting that differences in precision between isolated and clustered memory items were not due to different memory resolutions in these two situations. Instead, we tested the hypothesis that these differences occurred as a result of memory biases caused by neighboring memories, and we contrasted 3 different models (*Mixture-model fitting* in

Methods): an *attraction model* where responses to the target stimulus experienced a mean bias towards the neighboring memory; a *swap model*, in which responses to target stimuli were unbiased, but in some trials responses clustered around the neighboring non-target item; and an *attraction+swap model*, which combined the two situations: a fraction of swap responses and a mean bias toward neighboring memories (Figure 4.1.4a). Note that for the swap models we only considered swaps between close by items, which favors this model. We compared the estimated likelihoods of each model using Akaike information criterion (AIC). We compared the models by calculating the differences between AIC values (*Mixture-model fitting* in *Methods*). We calculated this difference between all the models and the best model. The best model (the one with the lowest AIC) was the attraction *model* for all but one participant, for which the attraction+*swap model* had the lowest AIC (ΔAIC for the *swap model* was 11.7, i.e. a relative likelihood < 0.0001). We excluded this subject to calculate the average information loss of the *swap* and *attraction+swap* models relative to the *attraction model* for the other participants. The swap model was the worst of the three statistical models tested (Figure 4.1.4b). Adding up AICs for these 8 participants, the relative likelihood of the *swap model* compared to the *attraction model* was below 10^{-4}. These results lead us to discard an explanation based on swap errors alone for the memory attraction.

To sum up, with this study we validated two behavioral predictions of the bump-attractor model, when storing multiple items. In particular, we found that humans have repulsive and attractive biases, as predicted by the model. Moreover, we discard that these attraction biases are driven by swap errors alone, a behavioral bias that occurs more often with high-load conditions, in contrast to our load 3. In fact, the model used here is not capable of simulating feature-binding, a fundamental element of this task that requires memorizing locations, colors, and their pairwise associations. This motivated us to extend our study in this direction, as described in the following section.

Feature-binding in working memory through neuronal synchronization[26]

Summary

Binding (or swap) errors occur in working memory tasks when a wrong response is in fact accurate relative to a non-target stimulus. These errors reflect the failure to maintain bundled in memory the conjunction of features that define one object, and the mechanisms implicated remain unknown. Here, we tested the mechanism of synchrony across feature-specific neural assemblies. We built a biophysical neural network model for working memory items defined by the combination of one color and one location. The model is composed of two one-dimensional attractor networks for working memory, one representing colors and the other one locations. Within each network, gamma-oscillations were induced during bump attractor activity through the interplay of fast recurrent excitation and slower feedback inhibition. As a result, different memorized items were held at different phases of the network's intrinsic oscillation. These two networks are then connected via weak cortico-cortical excitation, accomplishing binding between color and location through the synchronization of pairs of bumps across the two connected networks. In some simulations, swap errors arose: "color bumps" abruptly changed their phase relationship with "location bumps". Serial encoding of specific color-location associations was accomplished by stimulating briefly (50 ms), but strongly and simultaneously the corresponding bumps in each network. On the other hand, feature decoding was accomplished by stimulating the cued location with a .5 s pulse, which impacted strongly the corresponding phase-locked bump. Finally, the model makes specific predictions, testable at several levels. Firstly, delay duration and stimulus distance modulate swap errors. Secondly, swap errors in the model are associated with a lower phase consistency of oscillatory activity in the delay period.

[26] All the simulations and analyses were performed by me, but the suggestion to use phase-preservation index came from the Sreenivasan lab (NYU Abu Dhabi), our collaborators, which are validating that prediction in an MEG experiment (data not shown here).

Introduction

In biophysically-constrained models (Wang 1999), the interplay between a fast recurrent excitation, followed by a slower inhibitory feedback results in oscillatory activity. In cortical networks this arrangement is easily attained: fast excitation supported by fast AMPAR channels, interacts strongly with slower feedback inhibition mediated by $GABA_A$ receptors. The dynamics leading to oscillations can be described as follows. When the excitatory contribution of AMPAR channels to recurrent connections is large, these fast synaptic inputs produce 'bursts' of activity, which are eventually silenced by a slower, feedback inhibition. Without activity, feedback inhibition eventually wears off and the new excitation cycle ensues. Early work, modelling 1-item working memory with biophysically-constrained continuous attractor network models (Compte et al. 2000), has shown that embedding this excitation-inhibition interplay in ring-attractor models leads to oscillatory activity bumps during the memory period. Furthermore, these models can be adapted to store multiple memories (Edin et al. 2009; Wei et al. 2012; Standage and Paré 2018) and have been demonstrated to explain behavioral (Almeida et al. 2015; Nassar et al. 2018) and neurophysiological characteristics (Edin et al. 2009) of humans engaged in high-load working memory tasks. However, how multiple oscillatory bumps coexist in the same attractor network and which dynamics are generated has not been described, yet. Here, we advance in that direction and exploit these bump oscillation dynamics to model feature-binding in working memory. In particular, we focus on the simulation of behavioral biases originated from binding failures, also called *swap errors*.

Working memory load modulates oscillation power and frequency

We built a computational network model of a local neocortical circuit, with excitatory and inhibitory spiking neurons *(leaky integrate-and-fire* neuron model) connected reciprocally via excitatory AMPAR-mediated and NMDAR-mediated synapses and inhibitory $GABA_AR$ synapses (see Methods). The network model was tuned to support bump attractor dynamics with 3 simultaneous bumps (Edin et al. 2009), and further tuning of the relative weights of AMPAR and NMDAR-mediated currents set active reverberant neurons in the oscillatory regime. Using this computational model we started by investigating which dynamics originated within each network.

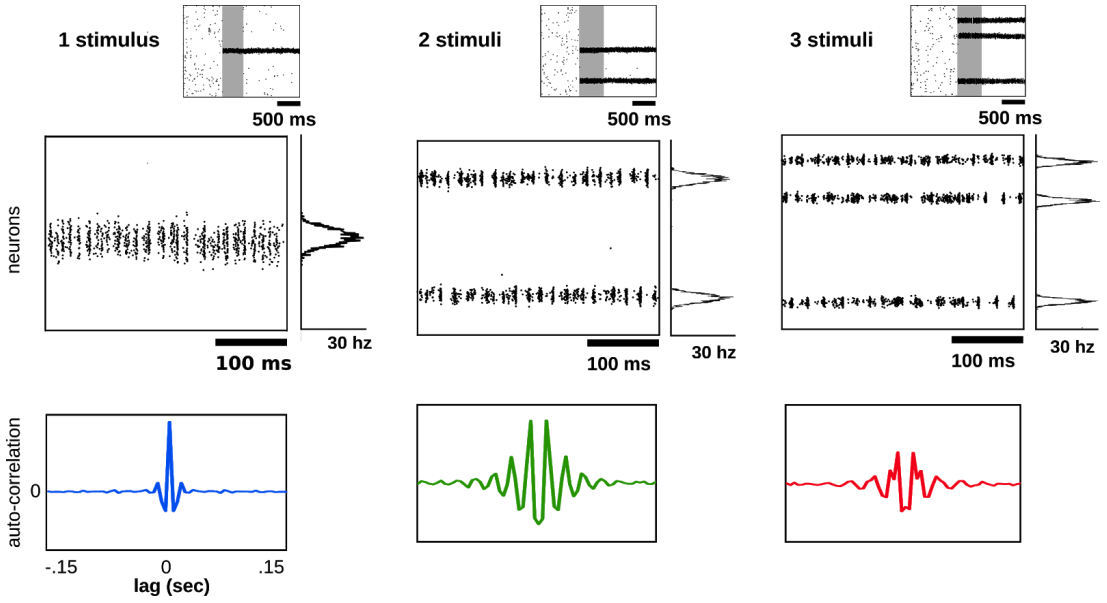


Figure 4.1.5. *Multiple bumps are spontaneously anti-correlated.* Top row, raster plots of 3 example simulations of load 1, 2 and 3. Middle. Zoomed versions of simulations on the top show clear oscillatory activity, confirmed by cross-correlation functions (bottom). For the load 1 case, we computed the autocorrelation. Notably, irregular activity due to external noise coexists with markedly oscillatory dynamics.

In our model, multiple bumps show anti-correlated oscillatory activity (Figure 4.1.5). As we store more bumps in the network, lateral inhibition originating from simultaneous memories establishes anti-phase dynamics during the memory period. Moreover, we found that the anti-phase behavior was robust in a wide range of values for AMPAR recurrent conductances (Figure 4.1.6).

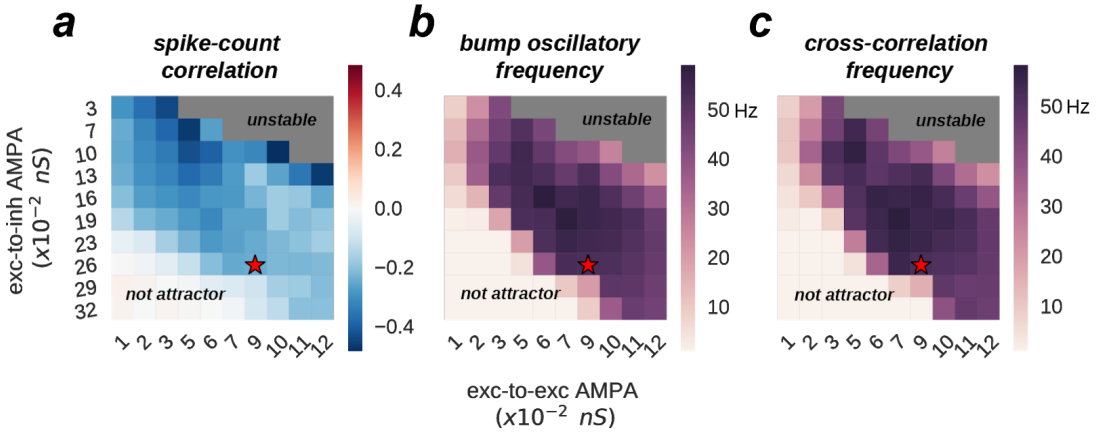


Figure 4.1.6. *Anti-correlated oscillatory dynamics as a function of excitatory recurrence (AMPAR conductance) in simulations with load 2.* **a)** Anti-phase dynamics as measured by spike count correlation between bumps. **b)** Frequency of the power spectrum peak computed using each bump's activity individually. **c)** Same as b) but computing the power spectrum of the cross-correlation between the two bumps (Figure 4.1.5, bottom). Red stars mark the parameter values of model simulations used throughout the study. This plot summarizes the dynamics of ~10,000 simulations (total) of 100 different networks.

Having seen this anti-phase dynamics between simultaneous bumps, we sought to contrast two alternative (and extreme) scenarios as we increase the number of stored memories (*memory load*). Under one alternative, bumps oscillate at a fixed frequency irrespective of load, so that the global network oscillation (adding up the activity of fixed-frequency out-of-phase bumps) would have a frequency that should increase linearly with memory load (scenario 1, dashed line Figure 4.1.7c). Alternatively, the network global oscillation could have a fixed frequency for different loads, and simultaneous bumps would take turns to fire in the available active periods. This would lead to halving the bumps' oscillation frequency as we double the memory load (scenario 2, dashed line in Figure 4.1.7d). We tested our model simulations to identify if our biophysical model adhered to one of these scenarios. To this end we ran multiple simulations with 3 different loads (presenting 1, 2 and 3 separate bumps during the encoding cue period) and we computed power spectra from either the aggregate activity of the whole network (network power) or from separate populations centered around each presented target (bump power). We then extracted the frequency of the peak network and bump power to study their dependency with load. We found signatures of both scenarios (Figure 4.1.7a,b). As we increase the memory load, the overall network activity oscillates at slightly increasing frequencies (Figure 4.1.7a,c). In contrast, each bump, corresponding to different memories, oscillates at markedly slower frequencies as load increases (Figure 4.1.7b,d). We quantified which were the dominant dynamics by plotting both the network's and each bump oscillating frequency against memory load. For better comparison, we normalized the frequency associated with different loads to the one of load 1. Moreover, we compared the effect of memory load against the aforementioned alternative scenarios (dashed lines in Figure 4.1.7c,d) and found that our network dynamics was more consistent with scenario 2. We therefore conclude that our biophysical network maintains a relatively constant global oscillation as more items are loaded into memory, and individual memory oscillations instead start skipping cycles to sustain out-of-phase dynamics with other memories.

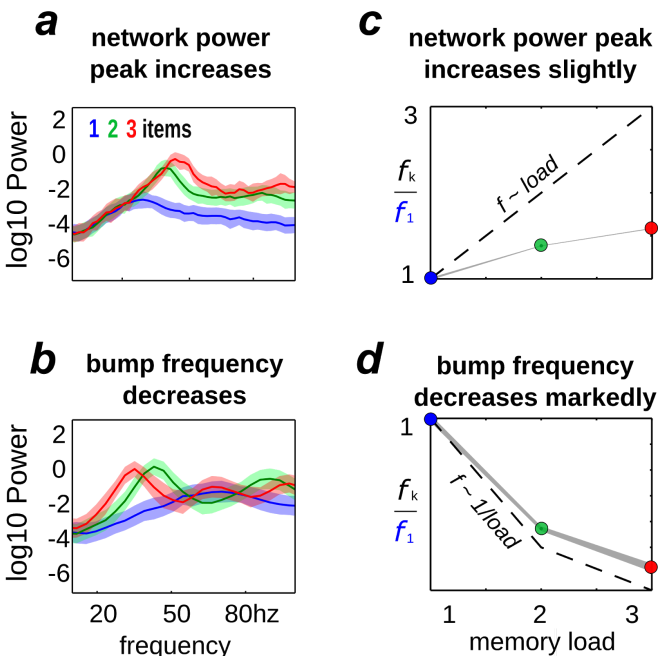


Figure 4.1.7. *Load-modulation of network and bump oscillatory dynamics.* Power spectrum computed from simulations of increasing load (1-3) using the activity of the whole network **a)** or of each bump's activity, **b)**. **c, d)** Peak-frequency computed from simulations with increasing load (f_k), normalized to simulations with a single bump (load 1, f_1) when computed from the whole network activity c) and only from each bump's activity d).

Thus, as shown before (Wang 1999; Compte et al. 2000), the interplay between recurrent (fast) excitation and (slower) feedback inhibition acting locally is the basis of the bump oscillatory behavior. Moreover, we now show that anti-phase dynamics of simultaneous bumps occurs due to bump competition, accomplished by lateral inhibition. Intuitively, this competition increases with memory load, leading to longer periods of silence during the delay-activity of each bump.

Uniform coupling achieves feature binding

How the conjunctions of different visual features are kept in mind is a long standing question in cognitive neuroscience (Schneegans and P. Bays 2018) - the so-called *binding problem*. However, there is consolidating evidence that different features of complex objects are maintained in independent stores (Delvenne and Bruyer 2004; Olson and Jiang 2002; Parra et al. 2011; Xu 2002; Wheeler and Treisman 2002). This suggests that different ring-attractors could be storing independent features, say 1 ring representing and memorizing colors and another ring storing locations (Ma et al. 2014). However, how these networks should interact to accomplish feature-binding is unclear.

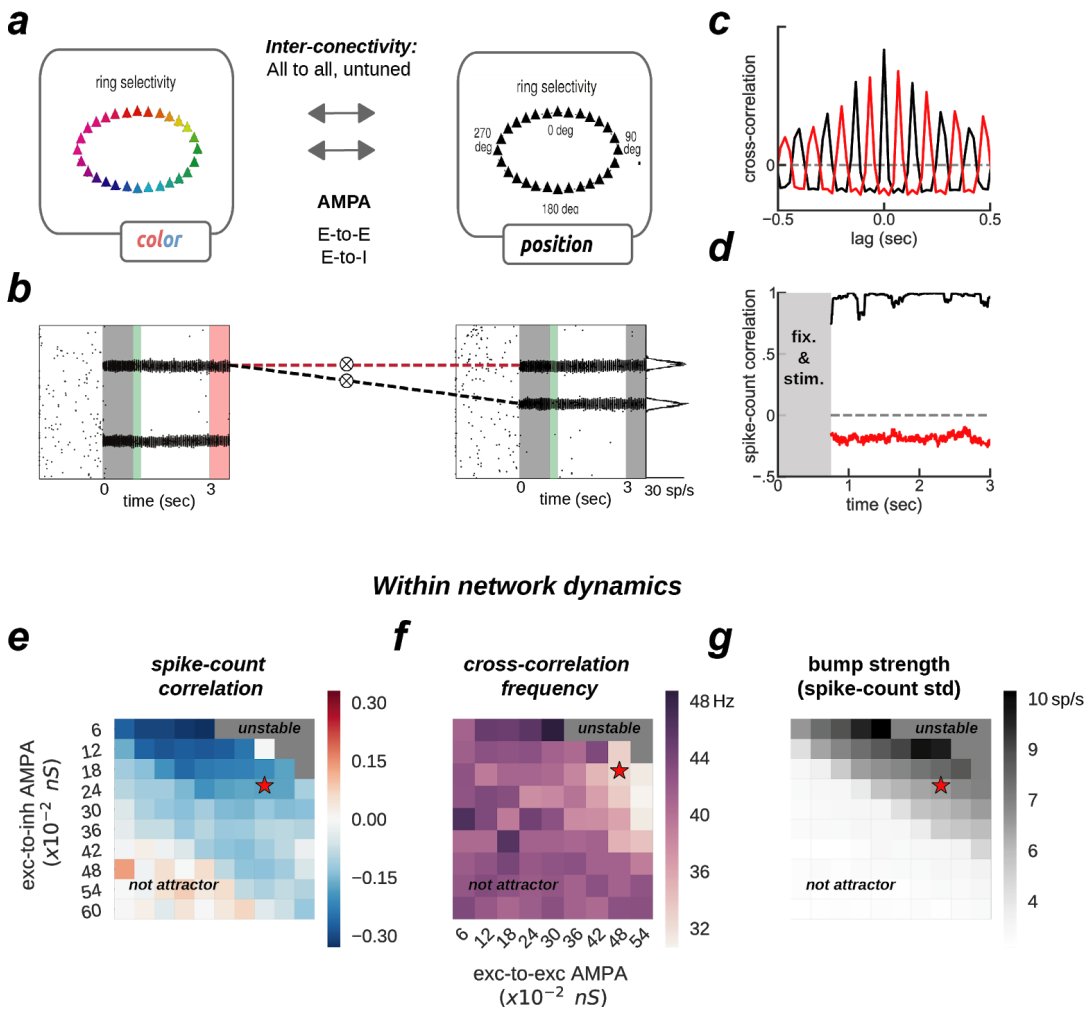


Figure 4.1.8. *Feature-binding through weak, uniform coupling.* **a)** Schematics of the final 2-network architecture, consisting of 2 ring-attractors with all-to-all, uniform connectivity. Each ring is able to store memories from one feature space (e.g. color or location) as activity-bumps (Figure 4.1.5). **b)** One example simulation for the two networks. The red-shaded area marks the period in which we read out the activity of the entire color network, while injecting current at one specific location in the location network (right gray-shaded area in the location rastergram, see main text for details about encoding/decoding). **c)** Cross-correlation computed between 2 pairs of bumps across networks (as marked with dashed red and black lines in panel b). For the black association, the cross-correlation peak is positive. In contrast, the cross-correlation peak was negative for the red association. **d)** Spike count correlation (in count bins of 5 ms, windows of 100 ms) of both associations through the memory delay is stable for this simulation. **e)** and **f)** same as Figure 4.1.6a,b, but for connected networks. e) Anti-phase dynamics within each network as measured by spike count correlation between bumps. f) Peak-frequency of power spectrum of the cross-correlation between the two bumps (Figure 4.1.5, bottom). **g)** Bump strength measured as spike-count variability at the end of the delay. e-g) summarize the dynamics of 22,000 simulations (total) of 100x2 networks.

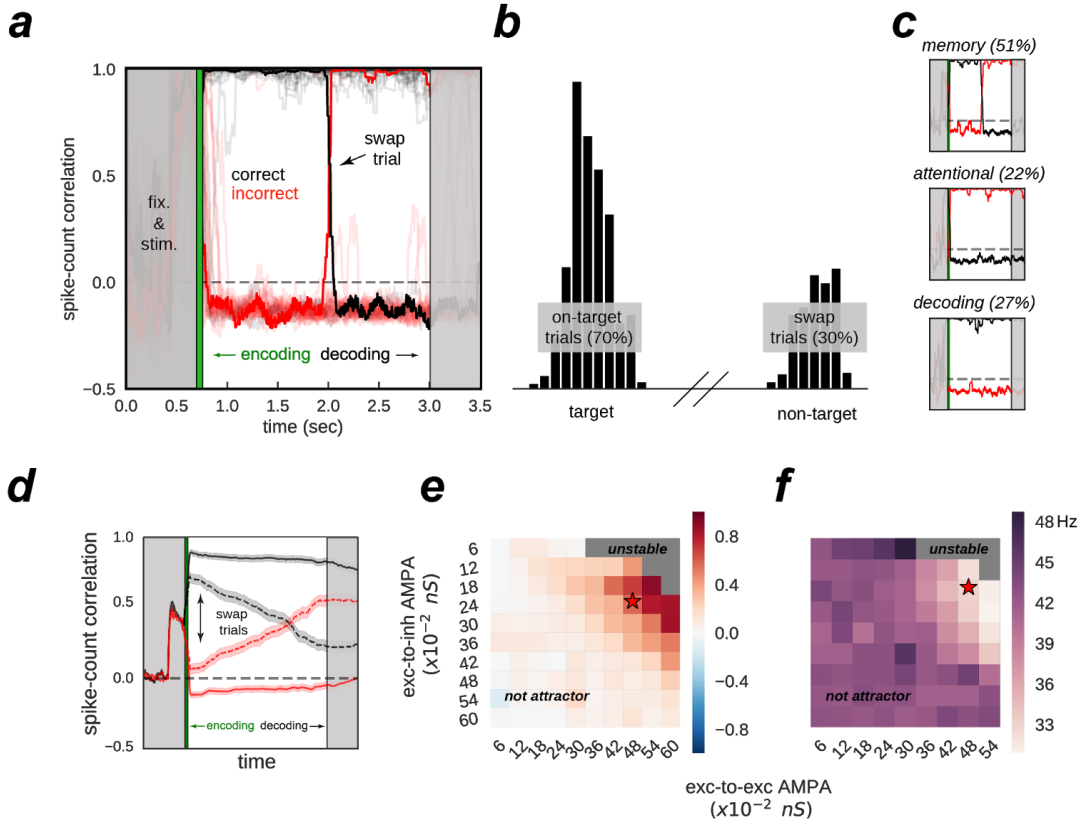


Figure 4.1.9. *Encoding and decoding is accomplished without temporal precision.* **a)** Spike-count correlation (in count bins of 5 ms and correlation windows of 100 ms) during the delay for 20 example simulations. During the encoding period (green), immediately after stimulus presentation, we bound two bumps, one from each network, by simultaneously stimulating them with external current. This ensured those two bumps were correlated through the trial more often than chance (correctly bound, black in the figures). On some trials (only one in a), noisy fluctuations changed the sign of this correlation suddenly (swap trial). During the decoding period (light gray, on the right) we simulated the cueing period of a real WM task, by injecting current on the cued location (0.5 s) of one network, while decoding mean firing rates from the other network. **b)** Histogram from 1,000 simulations. Bumps bound during encoding (target) were more likely to be read-out than unbound bumps (non-target). **c)** We identified three types of swaps, classified as memory swaps if the association changed abruptly during the delay (constituting 51% of the swap trials), attentional swaps if the wrong association was encoded (22%) or decoding swaps if the decoding fails (27%). **d)** Same as a), averaging across all trials separately for swap and on-target trials, as defined by the decoder, shown in b). **f, e)** summarize the dynamics of 22,000 simulations (total) of 100 connected (x2) networks as a function of inter-network connectivity. **f)** Binding measured as average spike-count correlation between correct bump pairs (bound pairs; red, in figures). **e)** Peak-frequency of power spectrum of the cross-correlation between bound bump pairs, across networks. Red star marks the parameters value of the model used throughout the study.

Here, binding between color and location is accomplished through the synchronization of pairs of bumps across two networks connected via weak cortico-cortical excitation (Figure 4.1.8). In particular, we connected two ring-attractors in the

regime described above with all-to-all, untuned excitatory connectivity. This connectivity was weak and it was mediated exclusively by AMPARs (Figure 4.1.8a), acting on all excitatory and inhibitory neurons. Interestingly, anti-phase dynamics within each network (as described above) was maintained robustly for a wide range of connectivity strength values (Figure 4.1.8, bottom row). Across networks, each bump's activity was in phase with one bump in the other network (Figure 4.1.8b,c, black) but out of phase with the other (Figure 4.1.8b,c, red). On the majority of the simulations, this selective synchronization was maintained through the whole delay period (see Figure 4.1.8c,d for an example simulation). This dynamics is therefore interesting as a possible mechanism to maintain bound the information of each presented stimulus. To this end, however, there are several aspects to resolve in relation to the encoding and decoding of this bound information.

On the one hand, synchronization selection was noise-induced in our simulations, resulting in across-networks associations between random pairs of bumps for different simulations. To control this association at the time of stimulus encoding, we stimulated strongly and simultaneously 1 bump in each network for a brief period of 50ms (Figure 4.1.8b, and Figure 4.1.9a, green period), forcing these 2 bumps (1 in each network) to engage in correlated activity during the delay period. Nevertheless, this phase-locked dynamics could be broken by noisy fluctuations, leading to possible misbinding of memorized features and swap trials (Figure 4.1.9a,b).

On the other hand, our model raised the question of how this binding of information could reasonably be decoded without resorting to complex mechanisms for spike coincidence detection. In our task the behavioral output, which consisted in answering which color was initially associated with a particular location, should depend on evaluating the pair of bumps in the 2 networks that maintained in-phase synchronization at the end of the delay. This was simulated as follows. For each trial, we probed one *location* by injecting external current to corresponding neurons in the *location network* at the end of the delay. This simulated the presentation of the location probe at the end of the delay (see Figure 1.1c in Introduction). This external current increased the firing rate in one of the location bumps, and we found that it also resulted in an increase of activity of the associated, in-phase *color bump*. Finally, we extracted the behavioral output with a maximum likelihood decoder applied on mean firing rate activity of the *color network* during the last .5 s, while probing the

corresponding location in the *location network* as described above. Figure 4.1.9b shows 1,000 of such simulated trials. Applying our encoding/decoding method to our simulations, results in 30% of trials wrongly associated with the non-target color (swap trials, Figure 4.1.9b). We then separated *swap* trials from *on-target* trials and computed the spike-count correlation in windows of 5 ms through the whole trial period (Figure 4.1.9d), and confirmed that on-target trials were in fact characterized by stable phase-locked activity, while the correlation between bumps in swap trials progressively approached the opposite dynamics (in-phase/anti-phase for the bound/unbound items, Figure 4.1.9d). Additionally, we identified three sources of swap errors in our simulations, classified as *memory swaps* if the correct association based on in-phase bump synchronization changed abruptly during the delay (constituting 51% of the swap trials), *attentional swaps* if the wrong association was encoded during the encoding period (22%) or *decoding swaps* if the correct association was encoded and maintained during the memory period, but the decoding failed (27%). See Figure 4.1.9c for example simulations.

Together, our biologically-constrained simulations demonstrate that feature-binding can be accomplished through selective synchronization. Crucially, encoding/decoding location-color associations was done without a *temporally precise code*, a long-standing limitation in the *binding by synchrony* framework (Shadlen and Movshon 1999). Moreover, we identified 3 sources of swap errors. Based on these computational findings, we investigated model predictions that could be compared with existing data or could generate hypotheses for new experimental studies.

Behavioral predictions: swap errors increase with delay (I) and item competition (II)

As discussed in the Introduction, swap errors have been described to depend on several task parameters. In particular, swap errors increase with delay duration (Pertzov et al. 2017) and decrease with target to non-target distances (Schneegans and Bays 2017a; Emrich and Ferber 2012). We therefore validated our feature-binding model against these behavioral findings. Firstly, Figure 4.1.10a shows that swap-errors increased with delay duration in the simulations. In our model, swap errors are induced by noisy fluctuations. Therefore, demanding longer delays will increase the probability of a large enough, swap-inducing noisy fluctuation. Secondly, Figure 4.1.10b shows how swap errors decrease with target to

non-target distances, congruent with previous findings (Pertzov et al. 2017). For very close locations, feedback inhibition is strongest, leading to “winner take all” dynamics between nearby bumps, explaining an increase of swap errors for such distances. For intermediate distances, similarly to Figure 4.1.1 (Almeida et al. 2015), simultaneous bumps interfere (repulsively and through their phase relationship, which is in this case less stable through the delay). Experimentally, these two regimes correspond to different scenarios. In the first case, one color is forgotten, while on the second scenario, there is an actual *swap* error. This prediction could be tested experimentally by probing the subject's memory on all items, instead of just one (Adam et al. 2017).

In sum, our model is able to describe a previously found dependence of swap errors with delay duration and with target to non-target distance, and it offers mechanistic explanations for such dependencies.

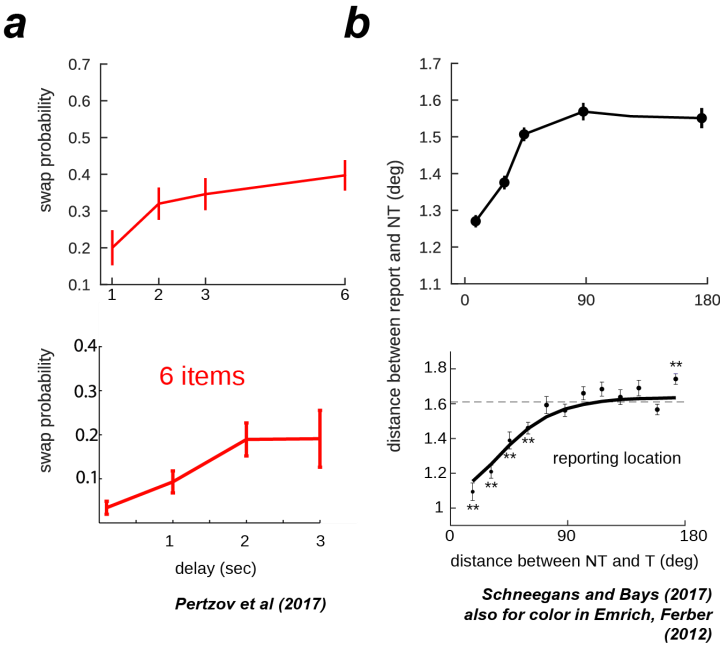


Figure 4.1.10. *Swap errors increase with delay duration and decrease with target-to-nontarget distances.* Model simulations (top) explain previous behavioral findings (bottom). **a)** Swap errors increase with delay duration and **b)** Simulations where target and non-target bumps are stored close-by increase swap errors, relative to when they are further apart.

Neural prediction: swap trials show less phase preservation through the delay

Finally, abrupt changes in the phase relationship between oscillating bumps is the central mechanism of swap errors in our model (Figures 4.1.9a,b). Therefore, it is worth deriving a testable prediction from this mechanism. Additionally, because these changes in phase relationships are abrupt, they require experiments using high sampling-rate techniques such as MEG or EEG, rather than the slower BOLD signal

that would smear out these events. Deriving a prediction testable using such techniques is therefore crucial.

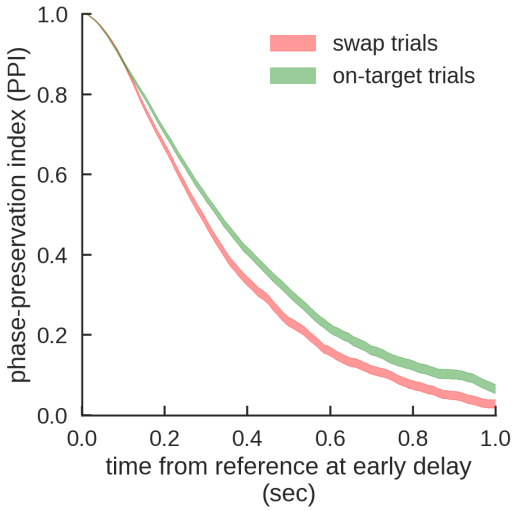


Figure 4.1.11. *Swap-trials show lower phase-preservation index.* Swap-error trials (red), compared with on-target trials (green) in the model are associated with a lower phase consistency of oscillatory activity in the delay period, as measured with phase-preservation index (PPI) using early delay as the reference time point.

Intuitively, swap errors in our model simulations are characterized by inconsistent phase relationships between brain signals when comparing the beginning and the end of the delay period. We therefore considered applying an analysis that has been proposed to test phase consistency in EEG/MEG: the phase-preservation index (PPI, (Mazaheri and Jensen 2006)). This was inspired through a collaboration with Prof. Kartik Sreenivasan (NYU Abu Dhabi, UAE). Accordingly, we first transformed our network's spiking activity in corresponding LFP's (Methods). We then calculated the phase-preservation index (PPI, see (Mazaheri and Jensen 2006) and Methods)[27] at the end of the delay, relative to the beginning of the delay, and separately for on-target and swap trials. As we expected based on our model simulations (Figure 4.1.9), this analysis applied to our simulated field data showed that trials containing swap errors had a lower PPI, compared to on-target trials (Figure 4.1.11). This specific prediction could be applied to MEG/EEG data recorded from humans performing this task, based on an analysis of behavioral responses able to discriminate swap error trials from correct and other error trials (Bays et al. 2009). This prediction is currently being tested in the Sreenivasan laboratory at NYUAD using MEG.

[27] Actually, testing our prediction with PPI was the idea of the Sreenivasan lab, our collaborators. In fact, this prediction is being tested in MEG experiments in the same lab.

Interim conclusions

We validated two behavioral predictions of the bump-attractor model, when storing multiple items. In particular, we found that humans have repulsive and attractive biases, as predicted by the model.

Furthermore, aiming to account for swap-errors, other sources of biases in multi-item working memory experiments, we extended the classical bump-attractor model. Our biologically-constrained model offers a plausible mechanism for feature-binding through selective synchronization. Importantly, it explains when this feature binding fails, including how it depends on delay duration and inter-item distances. Moreover, it provides a strong, testable prediction from its central underlying mechanism - phase-locked oscillatory activity during the memory periods.

4.2 Interference from previous memories

The interplay between bump-attractor and activity-silent dynamics in PFC underlies serial dependence in working memory[28]

Summary

Persistent firing of prefrontal neurons during maintenance periods of working memory tasks is a neural correlate of working memory. Alternatively, synaptic facilitation of recurrent connections has been proposed in theoretical models as a possible mechanism for working memory maintenance. Despite its theoretical appeal, there is little evidence supporting synaptic facilitation as the central mechanism of working memory. Instead, another class of models propose that it might play a complementary role in improving memory stabilization, with the cost of increasing history-dependent errors. We found several of these models' features in neurophysiological and behavioural experiments performed both in humans and monkeys. In particular, we found that old memories' information disappeared from spiking activity during the inter-trial interval (ITI) but reappeared prior to the forthcoming stimulus, as if it had been reignited from a hidden trace. In contrast, narrow cross-correlation peaks of simultaneously recorded neurons kept selectivity to the previous stimulus location through the whole ITI, supporting an extra, activity-silent memory system. A network model of bump-attractors with short term plasticity (STP) accounts for serial dependence and specifically explains these neurophysiological findings. Finally, the model predicts that reactivating old memories prior to forthcoming stimulus presentation should increase serial bias. We validated this prediction neurophisiolocially in human and monkey experiments by relating high decoding accuracy of previous-trial stimulus with an increase of serial biases.

[28] This chapter includes parts of a manuscript ready to be submitted: Joao Barbosa, Heike Stein, Rebecca Martinez, Adria Galan, Diego Lozano-Soldevilla, Kirsten Adams, Josep Valls, Christos Constantinidis and Albert Compte. *The interplay between bump-attractor and activity-silent dynamics in PFC underlies serial dependence in working memory.* Christos, Heike, Rebecca, Adria, Diego and Kirsten were involved in some aspect of the data collection. With the exception of the EEG data analyses, which were performed entirely by Heike, I performed all the simulations and analyses. Because I was involved in discussing the results during and after the EEG analyses, and because those results complement substantially our findings in the monkey PFC, I opted to include some of them here for completeness' sake.

Introduction

Working memory, the ability to maintain and manipulate information when no longer accessible to the senses, is considered a fundamental brain function in primates that underlies much of their enhanced cognitive capabilities. Understanding its mechanisms at the neural circuit level has thus been one major interest in contemporary systems neuroscience. The first proposed mechanism, supported by single-neuron recordings in non-human primates (Funahashi et al. 1989; Kubota and Niki 1971; Fuster and Alexander 1971), was selective sustained activity, whereby neurons in prefrontal and other cortices (Leavitt et al. 2017; Christophel et al. 2017) maintain an elevated firing rate during memory periods that is selective to the identity of the memorized item. Largely based on computational models, the initial cellular focus gave way to a primarily network mechanism in attractor networks supported by strong recurrent connections between neurons in the network (Compte et al. 2000; Wang 1999; Wang 2001). Specific experimental evidence in support of these network dynamics has been recently obtained in animal models during delayed response tasks (Wimmer et al. 2014; Inagaki et al. 2019). Recently, another mechanism has been proposed to support memory maintenance over short delays: stimulus triggered activity could charge some subthreshold mechanism (such as short-term synaptic plasticity) that would leave the network tagged with the identity of the previous stimulus without enacting it in neural activity (Mongillo et al. 2008). This latent memory trace would then be retrieved in spiking activity by means of a non-specific input to the network (Mongillo et al. 2008; Stokes 2015). This computational proposal has mostly received support from negative neuroimaging evidence: in some working memory tasks, even if memory performance is good, stimulus information cannot be retrieved from brain activity recorded in the delay period, but it is robustly decoded in other task periods. The apparent incompatibility of these two proposals (activity-based and activity-silent memory maintenance) has led to view them as alternative mechanisms, but modeling studies that have successfully implemented these activity-silent conditions invariably require the network to be configured close to the attractor network regime (Mongillo et al. 2008). Thus, these mechanisms rather than being regarded as alternatives may interact synergistically and support collectively different aspects of working memory function, as previously supported computationally (Barak and Tsodyks 2007; Hansel and Mato 2013). Here, we sought explicit evidence of such interaction with periods relying alternatively on either

mechanism in the course of a delayed-response spatial working memory task (Constantinidis et al. 2001a), by focusing on the encoding properties of brain activity during inter-trial periods. To this end, we capitalized on a behavioral read-out of such inter-trial dependencies that has recently captured much attention in the psychological literature. Serial biases in spatial working memory tasks denote small but systematic biases in reporting the location memorized in the current trial slightly attracted to locations memorized in the previous trial, especially when successive stimuli appeared in close proximity (Fischer and Whitney 2014; Papadimitriou et al. 2015; Fritsche et al. 2017; Bliss et al. 2017). Serial biases increase with memory delay length (Bliss et al. 2017; Papadimitriou et al. 2015; Fritsche et al. 2017), thus showing their dependence on memory maintenance. However, the mechanistic basis of serial dependence in working memory is still unclear, and both attractor dynamics and subthreshold activity-silent mechanisms have been proposed to carry stimulus-selective information from one trial to the next (Bliss and D'Esposito 2017; Papadimitriou et al. 2015; Kilpatrick 2018). In neural recordings from the frontal eye field (FEF) of monkeys performing an ODR task, (Papadimitriou et al. 2017) found neural firing selectivity to previous stimuli, prior to the new stimulus presentation. This finding is evidence that a persistent activity representation of the stimulus can remain in the circuit between trials, consistent with a maintained attractor mechanism. However, a critical component of their experimental design was a short inter-trial interval (ITI) of less than 400 ms. We hypothesized that longer ITIs would instead reveal a dynamic interplay between activity-based and activity-silent mnemonic network regimes that could be assessed neurophysiologically. Three main lines of evidence motivated this hypothesis: (1) dependencies of serial biases with delay and ITI durations are largely consistent with activity-silent, and not activity-based mechanisms (Kilpatrick 2018; Bliss and D'Esposito 2017; Papadimitriou et al. 2015); (2) the different quality of memory requirements in the delay and inter-trial periods of this task suggests different mechanistic substrates (Bliss et al. 2017; Papadimitriou et al. 2015); and (3) evidence for attractor dynamics in the network does not discard the activity-silent hypothesis, but instead suggests the network is close to the required regime for such dynamics (Mongillo et al. 2008).

Results

We trained four rhesus monkeys to perform an oculomotor delayed response task (ODR). The task consisted in remembering one out of eight possible spatial locations

at fixed eccentricity while maintaining fixation (Figure 4.2.1a). After a delay period of 3 s in which no stimulus was present, the monkey had to execute a saccade towards the remembered location. Responses were followed by a fixed inter-trial interval (ITI) of 2.1 s. In addition, we tested 35 human participants in variations of the ODR task performed by the monkeys, differing mostly in the report, which was done by a mouse click, in a variable delay duration (1 and 3 s) and variable ITI durations (between 1.1s and 5s, median 1.5 s). In all cases, we recorded the report location and computed behavioral errors as angular distances to corresponding target locations. Following the methods described in previous studies (Fischer and Whitney 2014), we analyzed the dependence of the current trial error on relative previous trial location. Despite substantial inter-species differences, both monkeys and humans showed a bias relative to previously remembered locations. This bias was attractive for short distances between previous-trial and current-trial locations, and repulsive for large previous-current distances (Figure 4.2.1a, 4.2.4a). We focused our research on specifying the neural mechanisms of between-trial memory underlying serial biases in this task, and their interaction with the known attractor dynamics underlying within-trial memory maintenance (Wimmer et al. 2014). To this end, we investigated electrophysiological measurements in the idle periods between successive trials of the task, including behavioral response, and fixation periods prior to the appearance of a new stimulus.

Reactivation of previous memory information prior to new stimulus presentation

We collected single-unit responses from the dorsolateral PFC (dlPFC) of two monkeys while they performed the task. As previously shown (Wimmer et al. 2014; Constantinidis et al. 2001a; Compte et al. 2003; Constantinidis et al. 2002; Constantinidis and Goldman-Rakic 2002), a substantial fraction of neurons in this area showed tuned persistent delay activity during the mnemonic delay period of the task (n=206/822). Moreover, (Wimmer et al. 2014) demonstrated that these specific neurons are part of bump attractor dynamics characterizing the memory periods of this task. Based on our hypothesis that an interplay of activity-silent and attractor mechanisms would support serial biases, we decided to focus our analyses on this subgroup of prefrontal neurons, and we grouped them in simultaneously recorded ensembles for decoding analyses (total of n=94 neural ensembles).

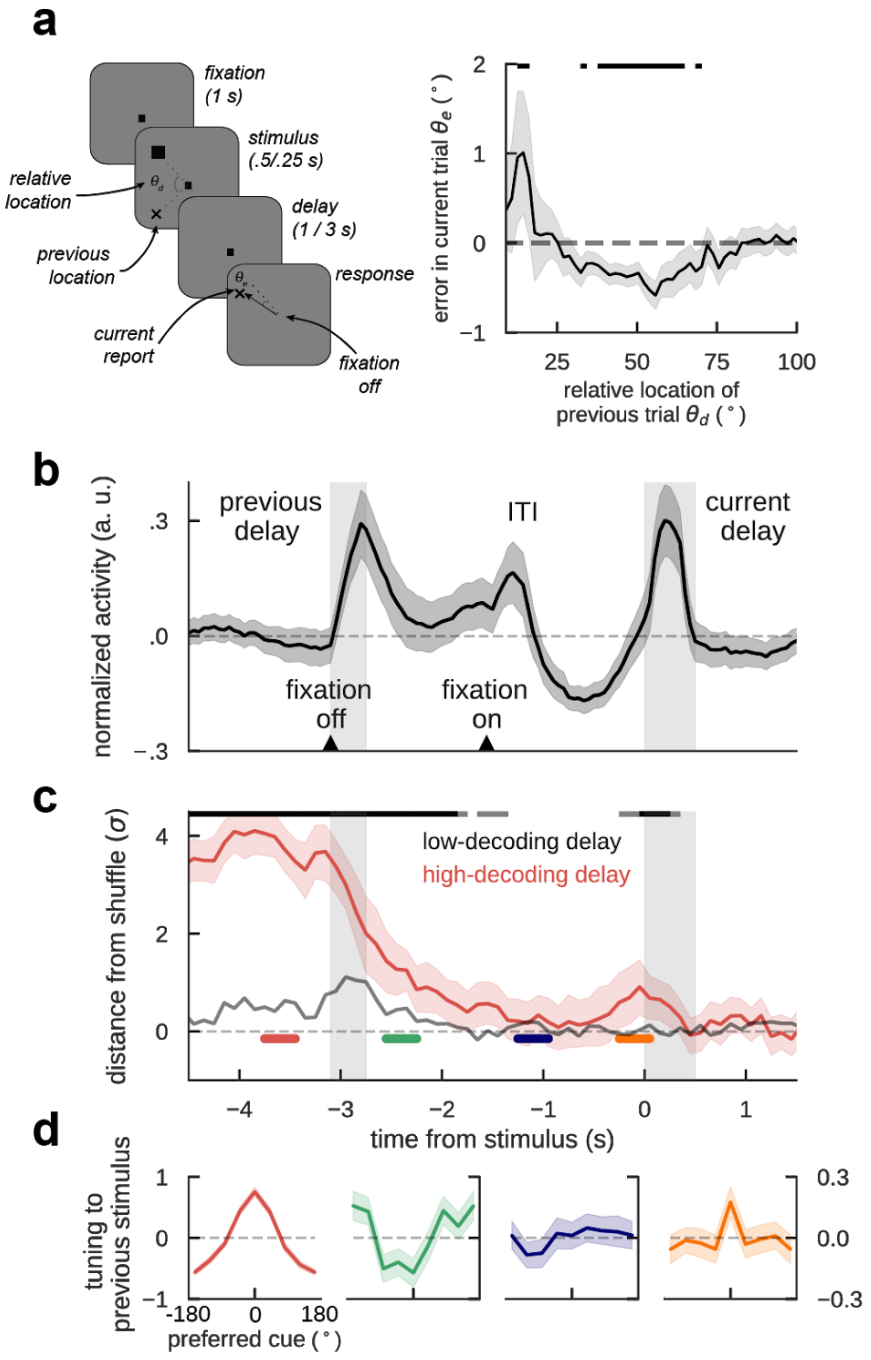


Figure 4.2.1. *Previous-trial stimulus code reactivates prior to the forthcoming stimulus*. **a)** Task design and error plot from 4 monkeys performing this task. Trials where the previous report was counter-clockwise to the current stimulus were collapsed into clockwise trials. Positive (negative) peaks mean that stimuli at that relative distance to previous-trial location elicited current-trial reports that were attracted to (repelled from) that location. Error bars are bootstrapped standard-error of the mean (SEM). Black solid bars represent p<0.05 computed using permutation test. **b)** averaged, normalized firing rate of n=206 recorded neurons during the intertrial interval. Left and right-handed gray background bars are response and stimulus presentation periods, respectively. **c)** Decoding accuracy of previous-trial stimulus, computed as distance from decoding of shuffled labels. Aligned with anticipatory ramping in pre-cue,

previous-trial stimulus reappears. Black bars on top mark timepoints for which decoding accuracy mean 99.5% C.I. is above zero; gray bars for 95%. **d)** Tuning to previous-trial stimulus computed at different epochs, marked in b) (p-value of one-sided t-test at prefered location: 9.85e-20 (red), 8.24e-3 (green), 0.43 (blue), 0.013 (orange), shadings are SEM). Unless stated otherwise, error-shading marks 95% C.I. of the mean.

Continuous recordings through a long ITI of 2.1 sec followed by a 1 sec fixation period showed that dlPFC single neuron average firing rates exhibited strong dynamics, compared to the stability during mnemonic delay periods (Figure 4.2.1b). Response execution and fixation onset were hallmarks in these dynamics, but we also noticed an increase of firing rate prior to stimulus presentation (Figure 4.2.1b), which could reflect an anticipation signal to upcoming stimulus presentation (because all ITIs in our experiment had a fixed duration, the monkeys could anticipate the forthcoming stimulus). We wondered if these changes in neuronal mean firing rate were also related to dynamical changes in stimulus selectivity. Under the attractor-based hypothesis for serial biases (Papadimitriou et al. 2017), sustained stimulus selectivity would be expected to extend from the previous trial delay period into the fixation period of the forthcoming trial. In order to access stimulus information during all consecutive trials, we trained a linear decoder on spike counts of small neuronal ensembles (n=94) of 1-6 simultaneously recorded neurons (Figure 4.2.2a). To test if each neural ensemble carried stimulus information, we attempted to decode stimulus location from 1000 surrogate datasets, consisting of label shuffles of the original dataset (i.e., containing no stimulus information by construction). This gave us a baseline distribution of decoder accuracy to which we could compare the decoding error on the original dataset (Methods). During the whole delay period, neuronal ensembles carried stimulus information and single neurons showed tuning to the corresponding stimulus (Figure 4.2.1c,d, red). Following the behavioral report, the memorized location was still decodable from ensemble activity, but plotting single neurons' tuning curves showed that at this time neurons had selective suppression of responses in their mnemonic preferred locations (Figure 4.2.1c,d, green). This could reflect neuronal adaptation mechanisms or else saccade preparation, as an anti-saccade was necessary to return to the fixation center. When single neurons recovered from this anti-tuning and were no longer tuned to the previous stimulus (Figure 4.2.1c,d, blue), decoding accuracy was not different from chance suggesting that information about the previous stimulus had disappeared in network activity. However, aligned with anticipatory ramping activity in the pre-cue period, previous stimulus information was again detected by the decoder just prior to the new stimulus

presentation, and single-neuron tuning reappeared (Figure 4.2.1c,d, orange). This is consistent with previous evidence demonstrating a firing-rate code for previous stimulus just prior to the cue (Papadimitriou et al. 2017), but it shows that the information is not continuously maintained through the ITI and fixation in dlPFC. Further, information reappearance occurred strongly (period marked in orange) in those neuronal ensembles that maintained more stimulus information during the delay period (Figure 4.2.1c and 4.2.2).

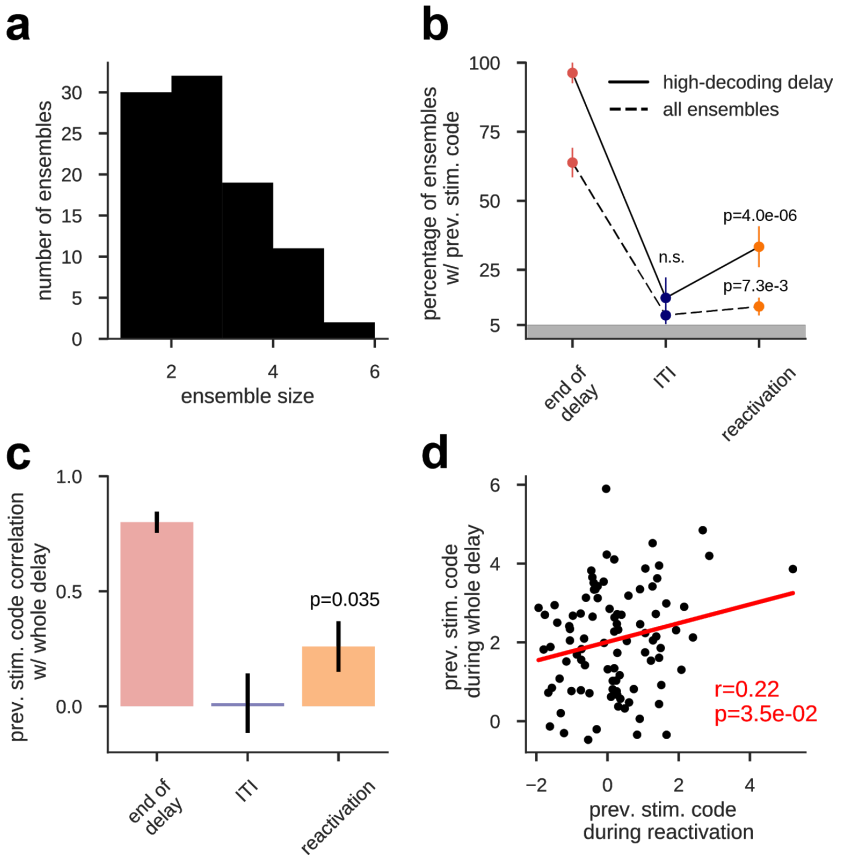


Figure 4.2.2. *The same neurons engaged in the mnemonic period are later reactivated.* **a)** Neuronal ensemble size simultaneously recorded varies between 1-6. **b)** fraction of neurons with previous stimulus code computed for all ensembles (dashed line) and only for the ensembles with highest previous stimulus code averaged across the whole delay. Selection of ensembles with highest delay code, reveals higher reactivated neurons. Statistical test done with binomial test at p=0.05 (n=96 and n=27, for all ensembles and highest delay code, respectively), standard errors are bootstrapped errors of the mean. **c)** correlation (r-value) between previous stimulus averaged across the whole delay and different time points (p-values: 6.5e-30, 0.87, 0.035) R-values and error-bars, standard error of the mean, computed using linear regression. **d)** Single ensemble values from c), orange.

This indicates a possible relationship between mechanisms of delay memory encoding, and mechanisms bridging the ITI to re-enact this information in dlPFC prior to the cue. Finally, PFC neurons exhibited negative noise correlations - a signature of a diffusing bump (Figure 4.2.3, (Wimmer et al. 2014)) - exclusively during reactivation. Taken together, these results suggest that previous trial memory information remains latent in the prefrontal cortex despite temporary absence of selective neuronal firing, and that this latent code can be reinstated in attractor-like neural activity in the network. Before further testing this idea, we first wondered if similar electrophysiological traces were observed in humans using EEG.

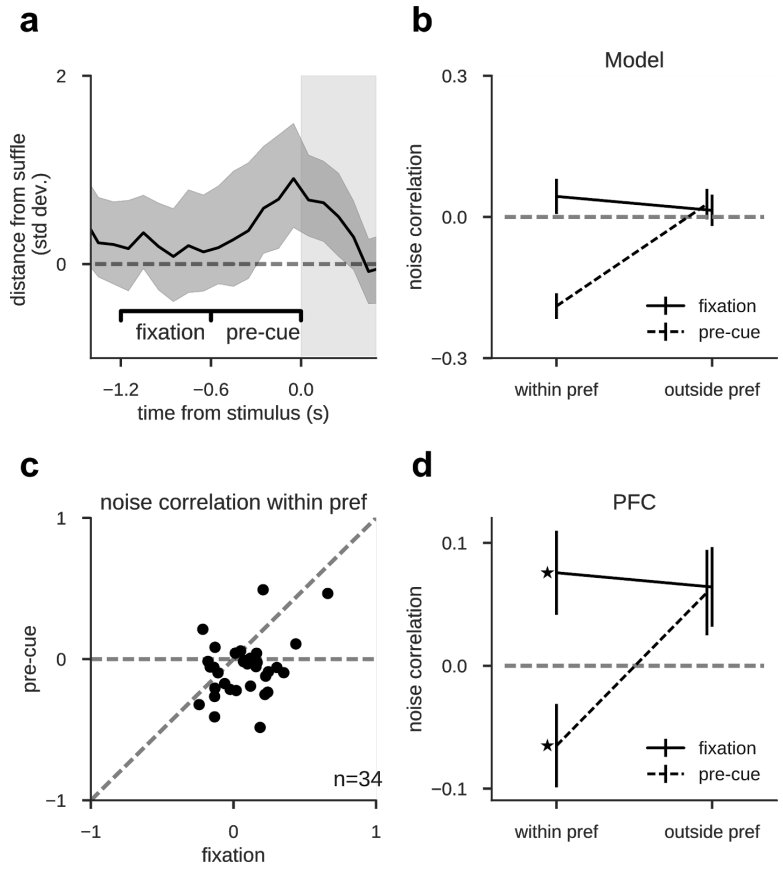


Figure 4.2.3. *Noise correlation between pairs of neurons is negative at reactivation as predicted by the model.* **a)** The model has different predictions for pre-cue and fixation period marked in this reproduction of Figure 4.2.1c decoding accuracies. **b)** Model simulations have negative correlations at pre-cue (reactivation period) only when conditioning on trials where the previous-trial stimulus was within prefered locations. **c)** noise-correlation of all simultaneously recorded pairs (n=34) for the within preferred condition. The shift towards the lower-right corner is consistent with the model predictions. **d)** Average noise-correlation of pairs in PFC exhibit same correlation pattern as predicted by the model (correlation negative during pre-cue (P=0.034, one-sided t-test), positive during fixation (P=0.018) and pre-cue different than fixation (P=0.0005) for the within preferred condition and non-negative during outside preferred condition.

Delay code is reactivated at fixation in EEG

We collected EEG from 43 scalp electrodes in 15 human participants, while they were performing the task. We extracted EEG alpha power from all electrodes and used a linear decoder with sinusoidal basis functions to predict the target location in each trial ((Foster et al. 2016), *Methods*). The target representation was significantly sustained during delay, response, and the next trial's pre-stimulus period (Figure 4.2.4b, diagonal axis). Considering that scalp EEG is a whole-brain measure, this code could be sustained by different representational components (e.g. stimulus, memory, response). We thus trained different linear decoders during delay (500-1000 ms after stimulus onset) around the response (250 before and 250 ms after), and used the respective weights to extract previous-stimulus information through different periods of the trial (Figure 4.2.4c). WM delay code is stable during stimulus presentation and delay, but disappears during response. Importantly, delay code reemerges after the next trial's fixation dot onset. In contrast, the response code did not generalize beyond the time at which the decoder was trained. We found reappearance of previous-trial stimulus tuning before the forthcoming stimulus (Figure 4.2.4c, right panel, and Figure 4.2.4d, lower panel), similar to monkey neurophysiology (Figure 4.2.1c), but at fixation onset instead of immediately before the forthcoming stimulus. Apart from showing a confirmatory correspondence with the time-course of mnemonic decoding in the monkey data, these human results also contribute additional insights. A whole-brain analysis failing to find continuity in the memory code through the ITI suggests that our finding of absent ITI rate code in the monkey's dlPFC was not due to being recorded from the wrong brain area, and instead was explained more parsimoniously by a latent memory code.

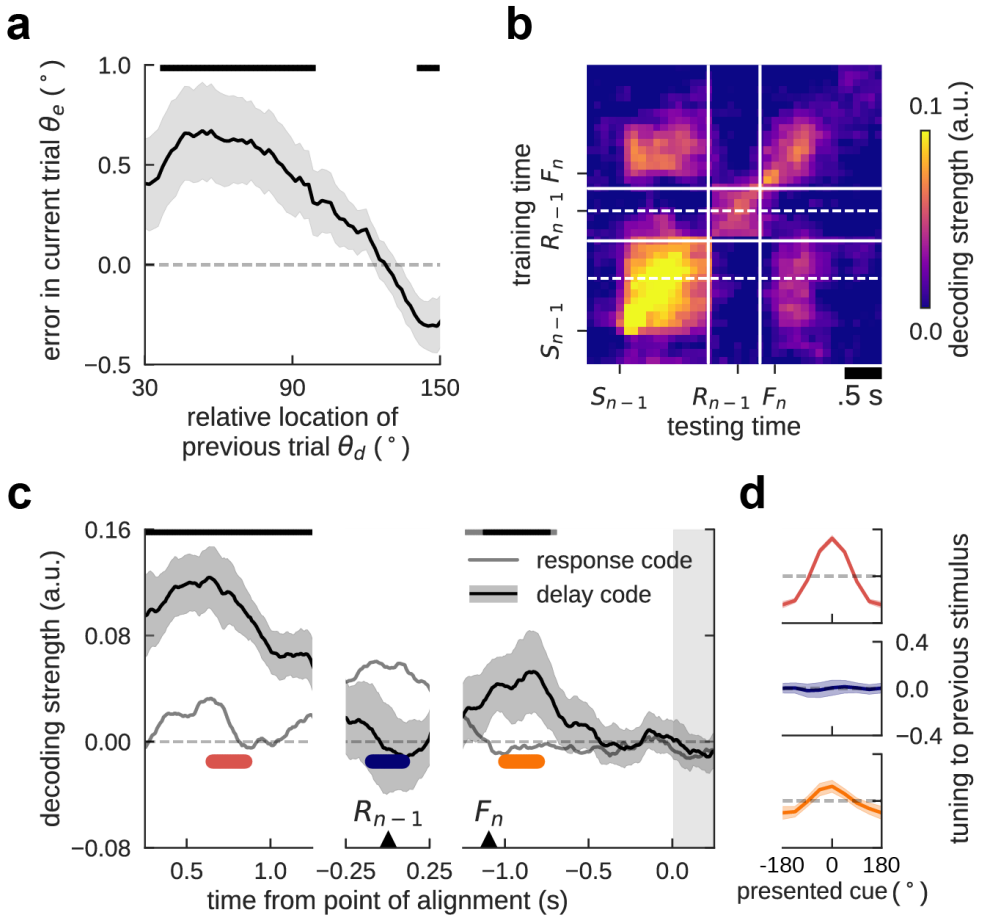


Figure 4.2.4. *Human EEG previous-trial delay code reactives at fixation onset.* **a)** Error plot for 15 human subjects performing the task. As is Figure 4.2.1 1, trials for which the previous location was counter-clockwise to the current stimulus were collapsed into clockwise trials. Error bars show one standard error of the mean (SEM), significance bars mark significant attraction (repulsion) with p<.05 (bootstrap test). **b)** Temporal generalization of previous-stimulus code for all combinations of training and testing time from previous stimulus onset (S_{n-1}), response (R_{n-1}), current trial fixation onset (F_n) to current-trial stimulus onset. Solid white lines mark the discontinuity of the EEG signal between delay, response, and current trial initiation. White dashed lines indicate the temporal center of the transversal selection for delay and response code shown in c). **c)** Tuning to previous-trial location during previous-trial delay (left), previous-trial response (middle), and current-trial fixation period (right). The cross-temporal decoder trained in previous-trial delay (.5s-1.0s) shows a stable delay code during delay, which disappears during the response period, and reappears prior to current-trial stimulus onset (black line with 95% C.I. error-shading), see also panel d). In contrast, the cross-temporal response decoder (trained -.25s-.25s after the previous-trial response) represents previous-trial information only during the response and stays around zero after current-trial fixation onset. Upper black bars show significant delay code decoding time points (99.5% C.I. of the mean above zero; gray bars for 95% C.I.). **d)** Average tuning reconstruction of the previous-trial stimulus at different epochs for the delay decoder, marked in c) (p-value of one-sided t-test at prefered location: 3.85e-9 (red), 0.92 (blue), 0.03 (orange), shading shows SEM).

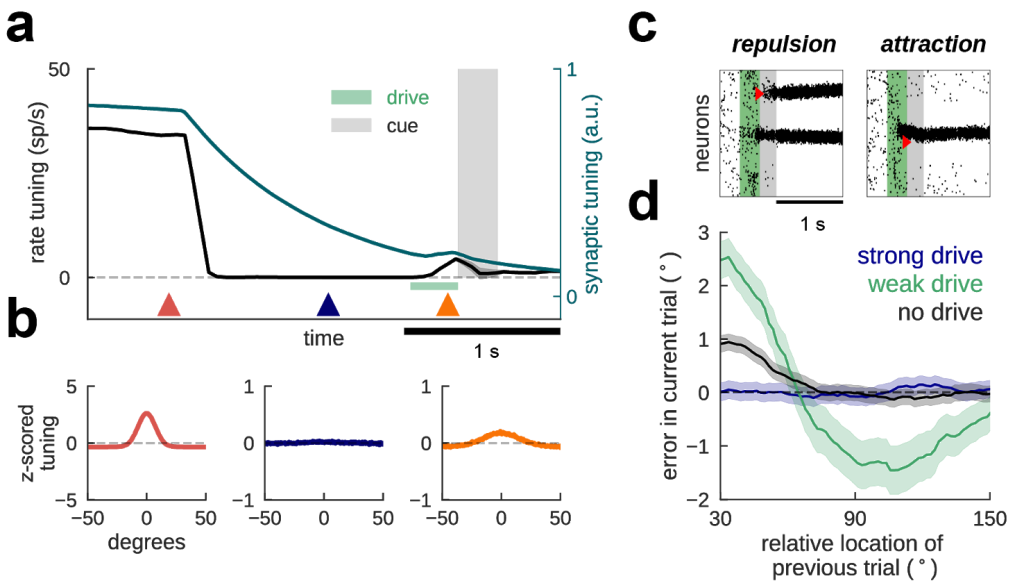


Figure 4.2.5. *Bump-attractor model with STP accounts for serial dependence and neurophysiology.* **a)** Average rate and synaptic tuning for all simulations in d). During the mnemonic period, both rate and synaptic tuning are at their maximum, both driven by neuronal activity sustaining a bump (red plot in b). At the end of each trial, a inhibitory input resets the network state, pushing it back to baseline activity (blue plot in b). This is reflected in the absence of rate tuning, but facilitated synapsis keep a latent bump that decays slowly enough to be reactivated by a non-specific stimulus marked in green (orange plot in b). **b)** Averaged single-neuron tuning to previous-trial stimulus at different epochs marked with colored triangles in a). **c)** Two example simulations, one depicting repulsion and another attraction from the previous stimulus location. Current stimulus marked with red triangle, while previous stimulus was at 180º for both simulations. Prior to the external stimulus, an anticipatory signal modeled as an excitatory drive to all neurons in the network (green) reactivates a bump from a latent trace in the previous-trial location, and interferes with the forthcoming stimulus repulsively or attractively, depending on the inter-stimuli distance. **d)** Behavioral responses computed from n=5000 for each curve. A weak anticipatory drive increases attractive serial biases and produces a negative lobe, while a strong drive removes serial biases.

Bump-reactivation as a mechanism for stimulus information reappearance

Based on our electrophysiology results and prior modeling studies (Mongillo et al. 2008), we formulated the bump-reactivation hypothesis to explain our data, and we implemented it in a computational model. We hypothesized that information held in memory as an activity bump during the previous trial's delay period (Wimmer et al. 2014) would be imprinted in neuronal synapses as a latent, activity-silent trace during the ITI. This latent bump could be reactivated due to the unspecific anticipatory signal, seen in mean firing activity in PFC (Figure 4.2.1b), or anticipatory mechanisms following an external cue predicting stimulus-presentation, such as the

onset of a fixation dot (Figure 4.2.4c). To test the bump-reactivation hypothesis, we built a bump attractor network model of spiking excitatory and inhibitory neurons with short-term plasticity (STP) dynamics in excitatory synapses (Methods). In each trial, stimulus information is maintained in activity bumps during the delay, supported by recurrent connectivity between neurons selective to the corresponding stimulus. During the whole ITI period, model neurons had no detectable tuning to the previous-trial stimulus (Figure 4.2.5a, black line and Figure 4.2.5b, blue) (Bliss and D'Esposito 2017; Kilpatrick 2018). However, the synapses of those neurons that had participated in memory maintenance in the previous delay were facilitated due to STP (Figure 4.2.5a, blue line). As a result, single neuron tuning could be recovered from the hidden synaptic trace using a nonspecific drive to the whole population (Figure 4.2.5a,c, *Methods*). Our computational model was thus an explicit implementation of the bump-reactivation hypothesis that we had formulated.

Serial biases can be explained by bump-attractors with short-term plasticity

We next used our computational model to derive behavioral and physiological predictions to test in our data, in particular in relation to serial biases. In order to simulate serial biases with our computational model, we ran 5000 pairs of consecutive trials with varying distance between the stimuli presented in each trial. We used the final location of the bump in the second trial (current-trial memory) as the "behavioral" output of the model in that trial. By applying parallel analyses to the behavioral experiments, we were able to model the profile of serial biases observed experimentally (Figure 4.2.5d, black), similar to (Kilpatrick 2018; Bliss and D'Esposito 2017). We then tested the impact of bump-reactivation in serial biases by comparing the behavioral output of simulations with and without anticipatory nonspecific drive before the second trial stimulus (Methods). We found that bump reactivation resulted in stronger attractive biases for similar successive stimuli, and in the appearance of repulsive biases for more dissimilar successive cues (Figure 4.2.5d, green line). We found that tuned intracortical inhibition was necessary for this emergence of repulsive biases upon bump reactivation (not shown). This showed that these repulsive biases are caused by repulsive interactions between simultaneously active bumps in the network ((Almeida et al. 2015; Nassar et al. 2018)), and are absent when there is no reignited bump that recruits localized inhibition at the flanks of the pre-cue bump of activity. We finally tested the dependence of this behavioral effect on the strength of the nonspecific drive. A very

short but strong impulse to the whole network during the ITI quickly saturated all the synaptic facilitation variables, effectively removing all serial biases in the output of the network (Figure 4.2.5d, blue). Thus, in this model bump reactivation affects serial biases non-linearly as reactivation strength is varied. In sum, our model can reproduce behavioral and neurophysiological findings described in Figure 4.2.1 and Figure 4.2.4 and derives several predictions that we tested experimentally.

Increased cross-correlation suggests a latent trace during ITI

Before addressing the serial bias predictions of the bump-reactivation hypothesis, we first sought an experimental validation that subthreshold mechanisms in dlPFC still maintained information about the previous stimulus during the ITI. The central peak of the cross correlation function is often used to assess the functional connectivity between pairs of neurons (Aertsen et al. 1989; Fujisawa et al. 2008; Amarasingham et al. 2012). We applied this analysis to the spike times of neurons in our computational model during the ITI, and then validated the prediction in the experimental data. A scheme in Figure 4.2.6a explains the analysis, which is as follows. First, we selected collinear pairs of neurons (distance between preferred locations *pref1* and *pref2* smaller than 8° in model and 30° in the data (n=67), Methods). We then compared the zero-lag peak of the jitter-corrected cross-correlation function (Methods) of all collinear pairs for two conditions: for trials in which the previous stimulus was shown close to their preferred locations (*pref*, maximum distance of 30° from previous location to either pref1 or pref2) or away (*anti-pref*, all other trials). The logic behind this analysis is that for trials in the pref condition, but not in the anti-pref condition, neurons had shown elevated activity in the previous trial and therefore their synapses would be facilitated. Indeed, our simulations supported our intuition and the ITI cross-correlation peak of pairs of excitatory model neurons maintained selectivity to the previous stimulus (Figure 4.2.6b) even if there was no selectivity in single neurons' firing rate (Figure 4.2.5a, blue). This cross-correlation selectivity with absent firing-rate selectivity reflected lingering synaptic traces sculpted in our model by the previous trial bump.

We then applied this method to the dlPFC data (Figure 4.2.6c) to assess if latent traces were still maintaining stimulus selectivity in the ITI. To this end, we constructed a cross-correlation selectivity index (CCSI) by subtracting the amplitude of the zero-lag peak of the cross-correlation for pref and anti-pref trials for each neuron pair (Figure 4.2.6b). We found that, when computing CCSI using data from a period in the

ITI where firing rate had ceased to represent the stimulus (Figure 4.2.1c,d, blue), CCSI was significantly positive (one-tailed t-test, t=3.11 p=0.001 n=67), reflecting selectivity to previous stimulus. However, we found that CCSI vanished in the reactivation period (Figure 4.2.1c, orange; one-tailed t-test, t=-0.29 p=0.39 n=67). This result violated the assumptions of our model, where synaptic selectivity should be maintained through fixation. We realized that synaptic enhancement of excitatory or inhibitory interactions could have cancelling effects in the CCSI and we decided to test them separately. We divided the selected pairs based on their cross-correlation peak sign in excitatory (*exc*) and inhibitory (*inh*) interactions, when their interaction had, respectively, an average positive or negative peak at zero-lag for both pref and anti-pref trials using spikes in the whole trial ([-4.5 s, 1 s]). We found a significant interaction between group (exc/inh) and time window (ITI/reactivation) for CCSI (two-way ANOVA, F=4.75, p=0.032, n=47). This indicated that cross-correlation selectivity for excitatory and inhibitory interactions was changing differently between these two different time points in the ITI. By testing them separately, we found that exc pairs had selectivity during the ITI (Figure 4.2.6d, orange) while inhibitory pairs represented the previous stimulus in the reactivation period (Figure 4.2.6d, green). This selectivity changed dynamically through the trial (Figure 4.2.6e): with the exception of immediately after report, where neurons showed anti-tuning to previous-trial stimulus (Figure 4.2.1c), CCSI for exc was always positive indicating stronger zero-lag cross-correlation when the previous stimulus was preferred (Figure 4.2.6e, orange). On the other hand, for inh interactions CCSI was negative (stronger inhibitory interactions following a preferred stimulus) only at pre-cue and previous-trial delay period (Figure 4.2.6e, green). This pattern is consistent with the latent memory mechanism residing in excitatory neurons and only being reflected in inhibitory interactions through the collective engagement in bump attractor dynamics, during the delay and at the time of reactivation. Thus, this cross-correlation analysis supports the hypothesis that previous, currently irrelevant stimulus information remains in prefrontal circuits in latent states, undetected by linear decoders that do not take precise spike timings into consideration.

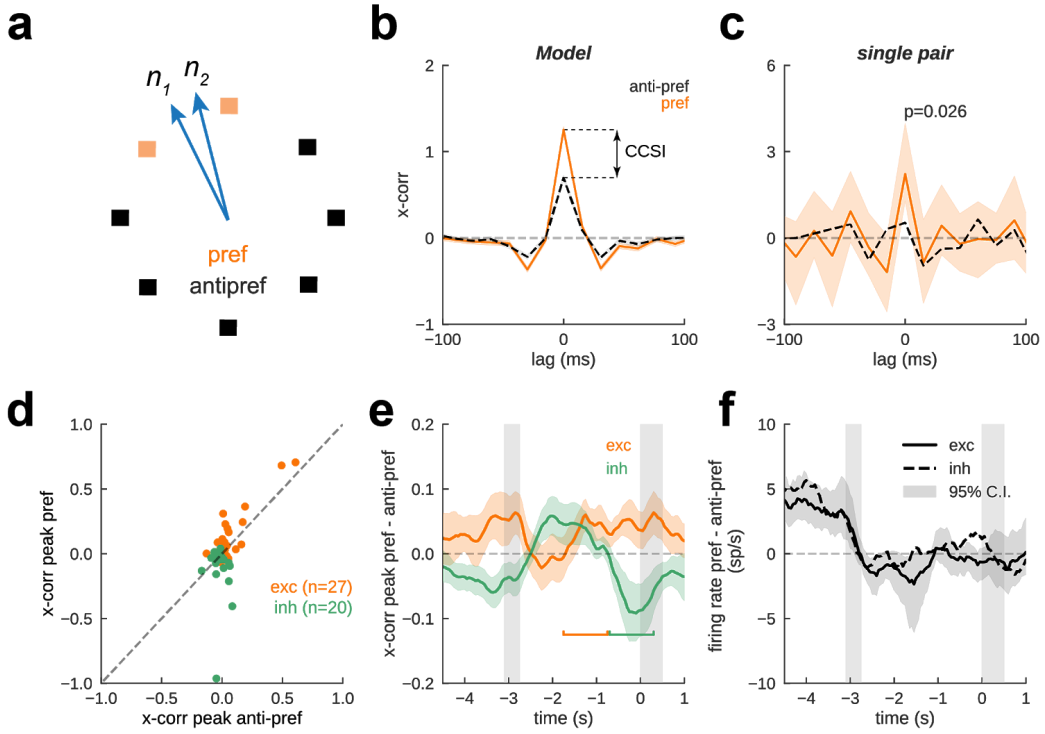


Figure 4.2.6. *Cross-correlation selectivity to previous-trial stimulus suggests an activity-silent trace in PFC.* **a)** Scheme of trial selection for the cross-correlation analysis. Pairs with similar prefered location (max 60°) and consistent peak sign during the whole ITI period - excitatory or inhibitory interaction pair (orange and green in d) and e) - were selected for the analysis (n=67). From all the memorized locations, we then separated those trials where the memorized stimulus location was close to the pair's preferred location from those that were far from the pair's preferred location (orange, pref and black, anti-pref, respectively). Methods for more details. **b)** The model predicts that, for each pair, spike count cross-correlation peak for pref trials (orange) should be higher than anti-pref trials (black), when computed during ITI, at which point there was no spiking selectivity (Figure 4.2.5a,b). We constructed a cross-correlation selectivity index (CCSI) by subtracting the amplitude of the zero-lag peak of the cross-correlation for pref and anti-pref trials for each neuron pair. **c)** Cross-correlation computed during 1 s period marked in orange in e) of an example pair recorded from the monkey PFC shows peak selectivity to previous-trial stimulus. Error-shading, C.I. of the mean. **d)** Same as c) but plotting cross-correlation peaks for all (n=27) exc pairs (orange, putative excitatory interaction) orange, and all (n=20) inh pairs (green, putative inhibitory interaction). Cross-correlation peaks significantly different using pref and anti-pref trials (one-sided permutation test, p=0.008, p=0.015) when computed during corresponding periods marked in e). Differences were still significant, after removing the 3 pairs with cross-correlation peaks 3 stddev above each group mean (p<0.03, both groups). **e)** Difference between cross-correlation peaks computed for pref and anti-pref trials, computed through the whole ITI period (1 s windows in steps of 50 ms) for putative excitatory (orange) and inhibitory (green) pairs. Except immediately after the report, where neurons show anti-tuning to previous-trial stimulus (Figure 4.2.1c), cross-correlation peak difference was positive for putative excitatory interactions, consistent with a latent, 'activity-silent' trace. On

the other hand, for putative inhibitory interactions, cross-correlation peak was negative only during previous delay and at pre-cue, consistent with the bump reactivation model. Colored bars mark time periods where cross-correlation was computed for d). Error-shading, bootstrapped standard error of the mean (SEM). **f)** Average firing rate difference between trials in pref and anti-pref conditions discards a potential confound between rate selectivity and cross-correlations peak selectivity. Error-shading are the bootstrap C.I. of the mean.

Previous stimulus reactivation increases serial biases

So far, our analyses have shown that the bump-reactivation hypothesis is consistent with neural activity recorded electrophysiologically, but we have not yet demonstrated a link between this activity and behavioral biases recorded experimentally to validate that re-activation of latent memory traces has the expected behavioral impact. We addressed this crucial point by designing analyses inspired in the behavioral predictions of our computational model (Figure 4.2.5). One prediction of the model is that the pre-cue reactivation of previous memories through population-wide anticipatory ramping of neural activity should lead to an increase in serial biases (Figure 4.2.5d). We tested this prediction in our neural recordings from monkey PFC and EEG recordings on the human scalp.

Monkey PFC. We first classified each trial based on leave-one-out decoding of previous stimulus in two different time windows during fixation: during a period with no stimulus information (ITI, Figure 4.2.1, blue), and just prior to cue presentation (reactivation period, Figure 4.2.1, and 4.2.7, orange). For each of these 2 windows we separated high-decoding trials (first quartile) from low-decoding trials (all other trials) and computed separately a measure of serial dependence (*Methods*) from behavioral responses. We found that serial biases were indistinguishable at ITI (Figure 4.2.7a) but they were stronger for high-decoding than for low-decoding trials at the time of bump reactivation (Figure 4.2.7b). This follows the prediction of our computational model, assigning behavioral relevance to the bump reactivation just prior to cue onset. We tested the robustness of our finding in two different ways. Firstly, we checked that this was not dependent on a singular selection of trial separations: for different separations of high- and low-decoding trials the serial bias strengths (*Methods*) changed smoothly and remained consistent with the reported result (Figure 4.2.8). Secondly, we also checked robustness of the result by testing not just two separate time windows in fixation but continuously through the trial: we repeated the same analysis, but classifying trials (low- vs. high-decoding) based on leave-one-out decoding computed at different time points of the trial. By calculating

the difference in serial bias strength (Methods) between low- and high-decoding trials, we found that a significant difference emerged only when trials were classified just before cue onset and serial biases remained virtually indistinguishable at all other time points (Figure 4.2.7c).

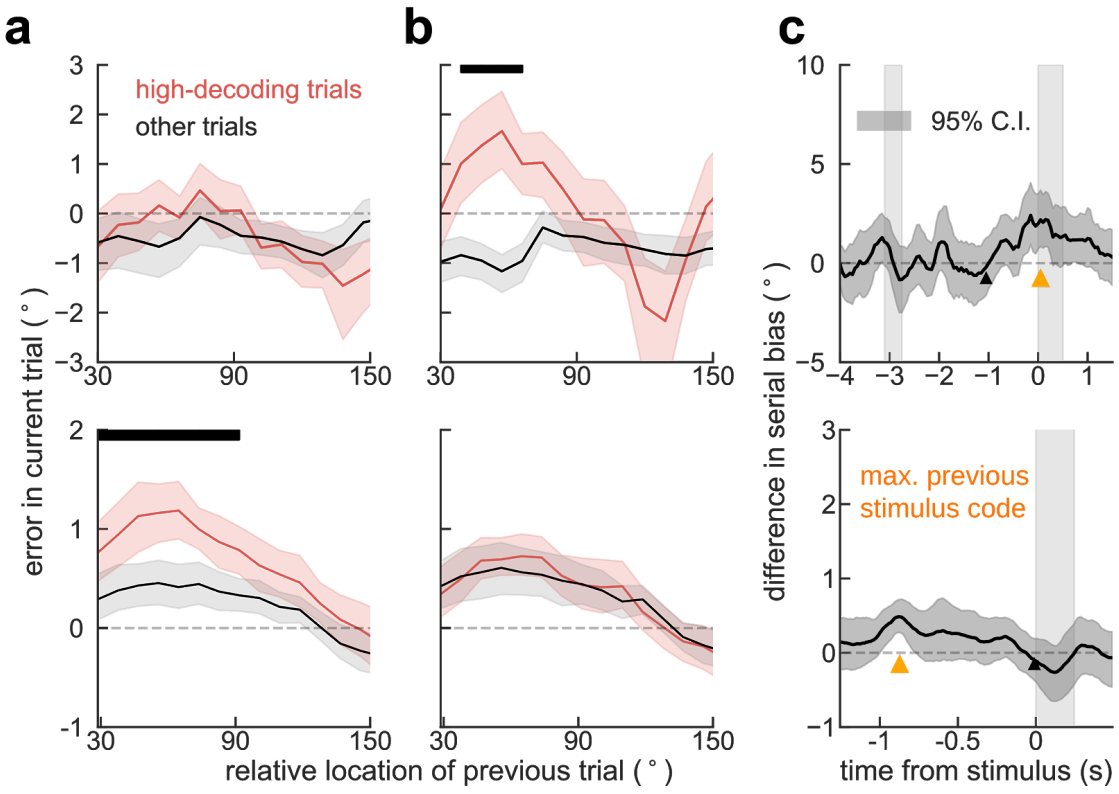


Figure 4.2.7. *Bump reactivation from a hidden trace increases serial biases.* Serial dependence plot separating trials where previous-trial stimulus information was high (red, higher quartile) and all other trials (black). In **a)**, a decoder was tested where there is no previous-stimulus information (before pre-cue, black triangle in e)), so high-decoding trials do not predict stronger serial biases. However, high-decoding trials have stronger attraction biases for locations close to previous-trial report, and repulsion for further distances when previous-trial stimulus information is estimated during pre-cue (**b**, orange triangle in e, the time point of maximum decoding in Figure 4.2.4c. **c)** Difference in serial biases (Methods) between high-decoding and other trials. During ITI, differences start to be significant at pre-cue, when a previous-trial bump is reactivated from a hidden trace. Triangles mark center of 1 s decoding windows for the two corresponding splits. **e,** and **f)**, same analyses as in a, b, and c), but for human EEG. Note that for humans, e) is the reactivation time point (orange triangle in g)), and f) the time point in which no stimulus information is present (black triangle). Black bars mark locations for which high-decoding trials had more serial biases than the other trials ($p<0.05$, permutation test). **g)** Difference in serial biases (Methods) between high-decoding and other trials. Differences are significant during fixation, at the time-point when previous-trial delay information reemerges in the cross-temporal decoder (Fiig. 2c). Triangles mark the center of decoding windows for the two corresponding splits in e) and f). In c) and g), time courses of differences between high-decoding and other trials are smoothed in time using a Gaussian filter.

Human EEG. Analogous to the analysis performed in monkey data, we grouped trials by their leave-one-out decoding accuracy. For each time point in the trial, we trained a decoder to decode the previous target position from the distribution of alpha-power across electrodes. We then computed the average decoding accuracy for the left-out trial in two different time windows: at reactivation period (Figure 4.2.4c,d and 4.2.7, orange) and just prior to cue onset, an arbitrary time point without stimulus information. As for the monkey data, serial bias was then calculated separately for high-decoding (top quartile) and low-decoding trials (all other trials). Consistent with our model prediction, we found stronger serial bias for high-decoding than for low-decoding trials for the reactivation period (Figure 4.2.7d), and not for the window at cue onset (Figure 4.2.7e), in which there was no reactivation of the memory code (Figure 4.2.4c). We further validated that this effect was specific to the point of delay code reemergence by repeating the same analyses for all other pre-stimulus time points (Figure 4.2.7f). Indeed, the prediction of behavioral bias was exclusively significant around the time of the delay-code reactivation (Figure 4.2.4c, orange). Taken together, these results support the hypothesis that previous trial memory reactivation prior to cue onset controls serial biases.

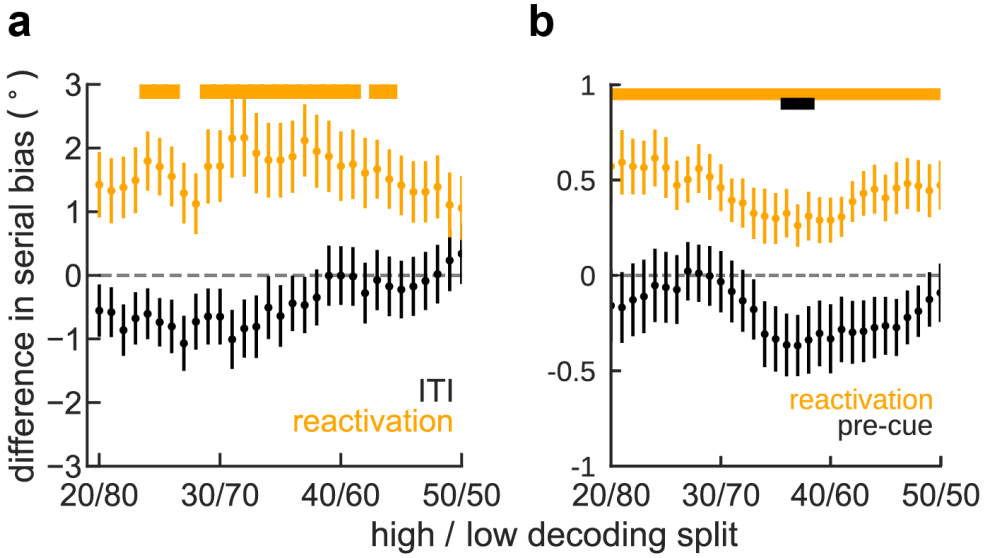


Figure 4.2.8. *The more trials are included in split between high- and low-decoding trials, the lower the serial bias* **a)** In monkey behavior **b)** In human behavior. Right after fixation, there is a negative correlation between the percentage of high-decoding trials included and absolute serial bias. In late fixation, when memories are not decodable anymore, there is no significant difference in serial bias for any percentage of trials included in the high-decoding subset. Error bars are 95% CI of the mean.

Build-up of serial biases in color working memory[29]

Summary

Serial dependence, how recent experiences bias our current estimations, has been described experimentally during delayed-estimation of many different visual features, with subjects tending to make estimates biased towards previous ones. It has been proposed that these attractive biases help perception stabilization in the face of correlated natural scene statistics as an adaptive mechanism, although this remains mostly theoretical. Color, which is strongly correlated in natural scenes, has never been studied with regard to its serial dependencies. Here, we found significant serial dependence in 7 out of 8 datasets with behavioral data of humans (total n=760) performing delayed-estimation of color with uncorrelated sequential stimuli. Moreover, serial dependence strength built up through the experimental session, suggesting metaplastic mechanisms operating at a slower time scale than previously proposed (e.g. short-term synaptic facilitation). Because, in contrast with natural scenes, stimuli were temporally uncorrelated, this build-up casts doubt on serial dependencies being an ongoing adaptation to the stable statistics of the environment.

[29]This includes parts of a study posted in biorXiv under the name “Build-up of serial biases in color working memory” by João Barbosa, and Albert Compte.

Introduction

Our perception depends on past experiences (de Lange et al. 2018). Serial dependence - how our current estimates are biased towards previous ones - has been described experimentally using many different paradigms (Kiyonaga et al. 2017; Bliss et al. 2017; Xia et al. 2015; Manassi et al. 2018; Czoschke et al. 2018; Alais et al. 2018; Manassi et al. 2017; Samaha et al. 2018; Suárez-Pinilla et al. 2018; Fischer and Whitney 2014; Liberman et al. 2016; Alexi et al. 2018; Cicchini et al. 2014; Fritsche et al. 2017; Taubert, Alais, et al. 2016; Taubert, Van der Burg, et al. 2016; Lieder et al. 2019). In particular, paradigms including delayed-estimations of different visual features (Kiyonaga et al. 2017), such as orientation (Fischer and Whitney 2014; Fritsche et al. 2017; Samaha et al. 2018), numerosity (Cicchini et al. 2017), location (Bliss et al. 2017; Papadimitriou et al. 2017; Papadimitriou et al. 2015), facial identity (Liberman et al. 2014) or body size (Alexi et al. 2018). It has been speculated that these ubiquitous attractive biases are a consequence of the world's tendency to be stable, and have the functional role of averaging internal noise (Cicchini et al. 2018; Kiyonaga et al. 2017; Fischer and Whitney 2014; Cicchini et al. 2014). Some have further argued that serial dependence is of adaptive nature, changing its strength depending on the stimuli statistics (Cicchini et al. 2018; Kiyonaga et al. 2017; Fischer and Whitney 2014; Taubert, Alais, et al. 2016; Cicchini et al. 2014). Color, which is strongly correlated in natural scenes (Cecchi et al. 2010), has never been studied with regard to its serial dependencies, possibly due to its strong systematic biases (Hardman et al. 2017; Bae et al. 2014; Panichello et al. 2018). Similar to other perceptual biases for other visual features (Gold et al. 2008; Sotiropoulos et al. 2011), these systematic color biases adapt to stimulus statistics in the course of one experiment (Panichello et al. 2018). This suggests that typical perceptual bias adaptations occur in time scales of minutes to hours. Slow adaptation of serial dependence, however, has never been characterized. If serial biases are also subject to adaptation with a similar time scale, when exposed to long sessions with uncorrelated stimulus statistics they should decrease or, in case of not being adaptive, they should remain stable. In fact, a recent study supports the latter hypothesis. In each trial, they asked humans to remember sounds which frequencies were sampled from uniform, gaussian or bimodal distributions. They too found that due to a working memory component, healthy humans had strong serial biases, but these were not affected by the stimuli distribution (Lieder et al. 2019) Here, we

address serial dependence in delayed-estimation color tasks, controlling for systematic biases and - contrary to our hypothesis - we characterize for the first time a slower dynamics of increasing serial dependence through the experimental session, despite uncorrelated stimulus statistics.

Serial dependence in color working memory

We analysed 8 datasets that are freely available online (Table 3.1, *Methods*), with a total of n=760 subjects performing variations of the same, delayed estimation of color task (Figure 4.2.9a). We will briefly describe the general experiment and Table 3.1 summarizes the specifics of each task, for detailed descriptions please refer to the original studies (Foster et al. 2017; Souza et al. 2014; Oberauer and Lin 2017; Bays et al. 2009; van den Berg et al. 2012). On each trial, a set of colored stimuli (varying from 1 to 8 stimuli) were briefly shown. After a delay period of roughly 1 second (see Table 3.1 for details), during which stimuli were no longer visible, subjects had to report the target color of a cued location. These color reports correspond to angles (i.e. degrees) on a color wheel rotated by a random amount on every trial, to avoid a spatial memory strategy.

We found that, across experiments, the subjects' reports were attracted to the previous target color for relative distances between previous and current trial target color of up to 90º in all experiments. Significant serial dependence occurred in all individual datasets for relative distances of up to 90º (Cam-Can, p=1.07e-09; Van der Berg I, p=0.006; Van der Berg II, p=9.66e-06; Oberauer & Li, p=0.007; Foster et al I, p=0.002, Foster et al II , p=4.35e-08; Bays et al, p=0.003), except for the dataset collected by Souza et al (Souza et al. 2014) (p=0.14).

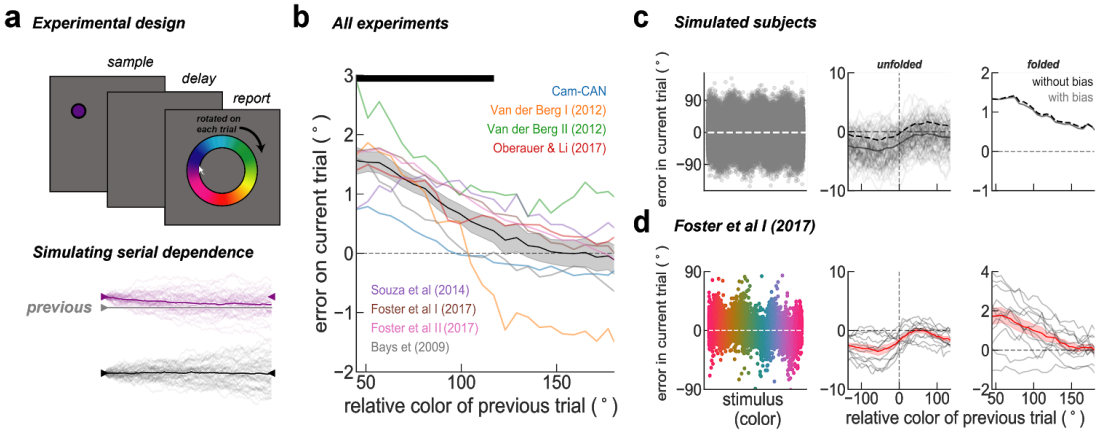


Figure 4.2.9. *Serial dependence in color.* **a)** Top, experimental design. All experiments were variations of a delayed-estimation of color task as in (Zhang and Luck 2008), differing mostly on set size and number of trials (Table 3.1). Subjects reported on a color wheel rotated by a random angle in each trial. Bottom, serial dependence was simulated as a drift towards the previous trial trace in a diffusion process. In purple, 50 trials with a stimulus feature (purple triangle) close to the previous trial trace (gray) and in black, 50 far trials. Thick lines represent the averages of each condition, which are attracted to previous trial stimulus for trials that are close by. **b)** Serial bias in the delayed-estimation of color task for all datasets. We found significant serial dependence relative to the previous report in all datasets ($p=0.0003$, t-test), except for the dataset collected by Souza et al (Souza et al. 2014) ($p=0.14$, t-test). **c)** Left, error to target stimulus reveals systematic biases on simulated trials. Middle, serial dependence calculated separately for trials simulated with and without systematic bias. Right, folded version of serial dependence removes all systematic biases without any additional preprocessing. **d)**, same as c) for trials of Foster et al I (Foster et al. 2017).

Folding serial bias curve removes systematic biases

Delayed-response reports are subject to systematic biases, which are particularly strong in the case of color (Bae et al. 2014). It has been argued that it is necessary to model and remove the systematic bias prior to estimating serial dependence (Bliss et al. 2017; Papadimitriou et al. 2015; Samaha et al. 2018). Here, we applied a model-free strategy that corrects serial dependence by "folding" the serial bias plot (Figure 4.2.9b). We tested this method in surrogate data obtained using a computational modeling approach. We simulated each delay of two consecutive trials as a diffusing memory trace (Wimmer et al. 2014) using a simple random walk simulation (see Method for details). On top of independent Gaussian errors responsible for diffusion, we added serial dependence as another source of error that accumulated incrementally at each time step, and two other sources of distortion (see Method for details): 1) systematic biases derived from inhomogeneities of the task space, and 2) systematic rotational biases (e.g. a constant clockwise error (Samaha et al. 2018; Fritsche et al. 2017; Bliss et al. 2017)). Figure 4.2.9c shows the effect of these systematic biases on serial bias estimation. We simulated $n=1000$ trials and $n=100$ subjects with (gray) and without (dashed black) systematic biases. As previously reported (Samaha et al. 2018), we found that systematic biases shift the serial bias function to negative values. This shift precludes the correct identification of attractive and repulsive serial bias regimes, and complicates comparison across subjects. We found that a simple processing of the data allowed for a model-free correction of systematic-bias-induced shifts: we "folded" the serial bias curve by collapsing all negative distances between consecutive targets on

positive values, while also inverting for these trials the sign of the behavioral error (*folded error,* see Method). This method effectively removes all systematic biases introduced in simulated trials (Figure 4.2.9c, right). For illustration purposes, we show the application of this method in one dataset (Foster et al I (Foster et al. 2017)) with similar systematic biases (Figure 4.2.9d, left), that led to a shifted serial bias function (Figure 4.2.9d, middle) and finally a *folded* version, without systematic biases (Figure 4.2.9e, right). Thus, our simulation approach validated the folding approach to correct for systematic biases in serial bias estimations, and allow for across-subject comparisons (Figure 4.2.9b).

Color serial dependence builds up in the course of an experimental session

Serial dependence, some argue, reflects the world's tendency to be stable (Cicchini et al. 2018; Cicchini et al. 2014; Kiyonaga et al. 2017). The reasoning is that because similar stimuli usually elicit similar behavior, the brain would incorporate mechanisms to exploit these patterns (Cicchini et al. 2018). Along these lines, a recent study has shown that systematic biases in color working memory change in the course of an experimental session to adapt to stimuli statistics (Panichello et al. 2018), arguing that systematic biases seen in delayed-estimation of color reflect real-world statistics. If similar adaptive plasticity operated for serial dependence in the time scale of the experimental session, we would expect to see a reduction of serial biases as one is exposed to a sequence of uncorrelated stimuli. To our knowledge, the stability of serial dependence within an experimental session is yet to be characterized. To address this question, we divided each session in two halves and computed serial dependence relative to the previous trial report for each subject and experiment in each of these two halves. When averaging all experiments together, we found that, contrary to our hypothesis, there was stronger serial dependence in the second half than in the first half of the experimental session (Figure 4.2.10a). To further characterize this serial dependence build-up, we used the folded error on each trial as a scalar measuring the evolution of serial dependence in the course of the session (see Methods). Figure 4.2.10b illustrates this analysis using a sliding window of 75 trials for Cam-Can (Shafto et al. 2014; Taylor et al. 2017) dataset and of 200 trials for the Foster et al I dataset (Foster et al. 2017), both showing a clear increase of serial dependence as the session progressed. To test this effect across subjects and experiments, we obtained the regression slope of the folded error as a function of trial number. We computed this slope for each subject and for all experiments. The

data shows that serial dependence build-up was positive for all but 1 dataset (Bays et al (Bays et al. 2009)), significant for 2 datasets individually (Cam-Can (Shafto et al. 2014; Taylor et al. 2017), p=0.008 and Foster et al I (Foster et al. 2017), p=0.0002, t-test) and for all combined (p=0.043, t-test on the 8 averages across experiments or lumping all subject together, p=0.006).

We proceed to test if the serial bias build-up was related with subjects getting familiar with the task, in which case one would expect to see an improvement in performance through the session, or related to subjects feeling tired, which should be reflected in worsening of performance. To this end, we calculated the fraction of guesses as a proxy of tiredness or engaging. We classified those trials with error > 90º in independent windows of 20 trials. Importantly, these trials were excluded from all the other analyses. We then computed the slope of change of the fraction of guesses, instead of the folded error as above. In fact, we found that subjects in 2 datasets significantly decreased their guess rate through the session (CamCan, p=7.1e-20 and Van der Berg I, p=0.008) and in other 2 increase their guess rate (Souza et al, p=0.02 and Bays et al, p=1.1e-06). However, this trend did not correlate with serial bias build-up for any dataset independently (p>0.35, linear regression), lumping all subjects together (p>0.35, linear regression) nor averaging across experiments (p>0.2, linear regression). Serial bias build-up was not correlated with subject's squared error either (p>0.45, lumping all subjects together). Together, these results show that serial dependence is not stable on the time scale of one experimental session, as previously assumed, and it also discards a mechanism that adapts to stimulus statistics. Instead, our result suggests the involvement of slowly accumulating plastic mechanisms in serial dependence of color delayed-estimations.

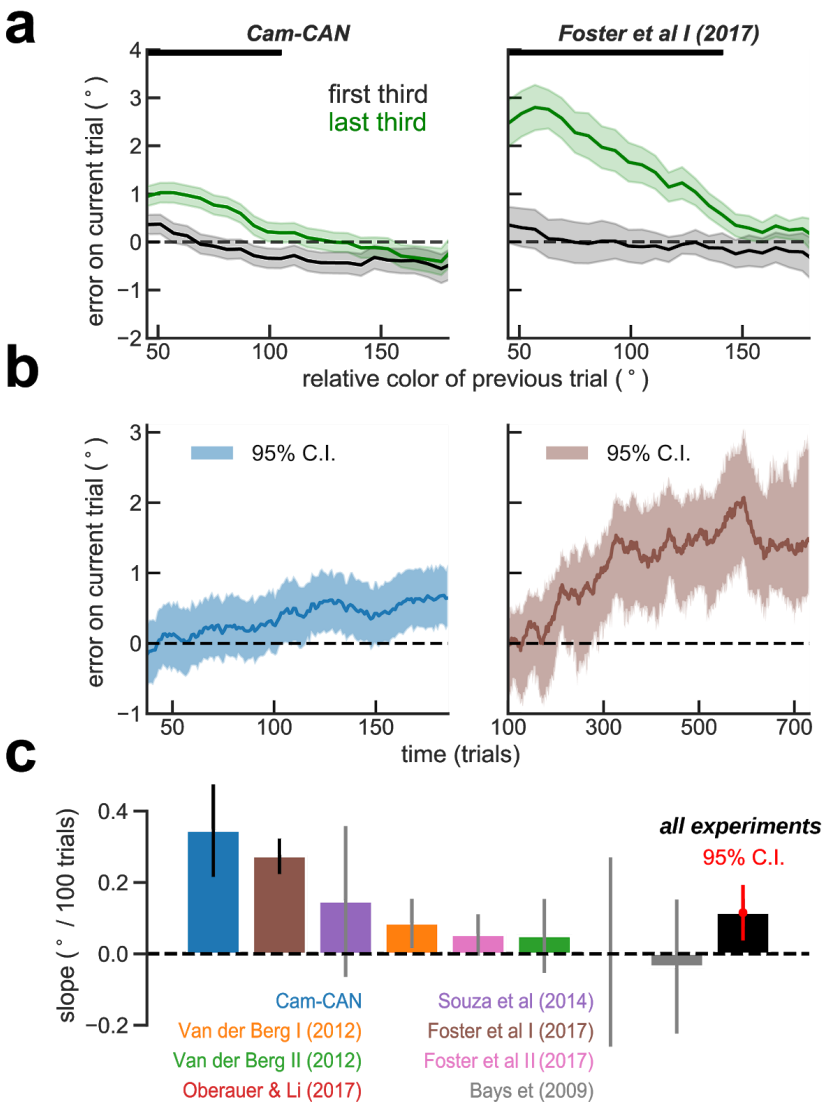


Figure 4.2.10. *Serial bias builds up during a session.* a) Serial biases computed using first third (black) and second third (green) of the trials for two example experiments: Cam-Can and Foster et al I (Foster et al. 2017). Black bars on the top mark where curves are significantly different, $p<0.05$, permutation test. b) Both experiments show a significant increase in serial dependence through the session computed with a sliding window of 75 trials for Cam-Can (Shafto et al. 2014; Taylor et al. 2017) and 200 for Foster et al I (Foster et al. 2017). c) For each subject, we computed the slope of serial dependence over the course of the session (without averaging). We found that seria-bias build-up was significant in two experiments and mark them with black error-bars. (Cam-Can, $p=0.008$; Foster et al I, $p=0.0002$). Error-bars were calculated from bootstrap distributions and unless stated otherwise, are S.E.M.

Interim conclusions

Combining humans and monkey neurophysiological experiments, we found that after memory-guided reports, past stimulus information ceased to be represented in PFC neurons. Surprisingly, this information was again represented by such neurons before furthcoming stimulus onset. Because this was reminiscent of 'activity-silent' mechanisms, we called this phenomenon *reactivation*. By introducing short-term plasticity dynamics in the bump-attractor model, we explained such reactivation dynamics and derived novel predictions that we then validated, linking behavior and neurophysiology. First, we found enhanced functional connectivity between PFC simultaneously recorded neurons that were involved in memory maintenanced, compared with trials that these same neurons were not involved in memory maintenanced. Second, in line with the simulations, we found that during trials undergoing higher information reactivation, monkeys and humans were prone to stronger serial biases. Together, these results point to an activity-silent mechanism, such as short-term plasticity, as the underlying mechanism of interference from previous memories in working memory (i.e. serial biases).

Finally, we described for the first time serial biases in color working memory. In particular, we found that attractive serial biases build up in the course of an experimental session, an effect that cannot be explained solely by short-term plasticity. Rather, to account for such effect, we tentatively speculate that metaplasticity mechanisms need to be included in existing theoretical models for working memory interference.

4.3 Reactivation of previous memories with non-specific stimuli

Reactivation of previous-trial memories with non-specific stimuli[30]

Summary

Recent studies assert that memories latent in activity-silent states, for example in facilitated synapses, can be reactivated by means of non-specific stimuli. To our knowledge, all these reactivation studies tested this hypothesis on EEG signals recorded from humans performing comparable working memory tasks, but differ on which type of non-specific stimulus was applied: a visual, full-field stimulus of 100 ms or a single-pulse transmagnetic stimulation (TMS). Alongside studies found that subjects performing equivalent working memory tasks exhibit small, but systematic biases towards previous-trial stimulus, the so-called serial biases. Besides, computational models propose that activity-silent mechanisms could underlie these biases. Here, we built one of those models and, inspired by the reactivation studies, we predicted that non-specific stimuli targeted to previously facilitated neurons should increase serial biases. Furthermore, we applied TMS or a visual stimulus during the inter-trial period of independent experiments, using the same single-item visuo-spatial working memory task and expected to see a modulation of serial biases in both conditions. In line with our computational hypothesis, serial biases were modulated when TMS was applied to the dorsolateral prefrontal cortex. However, we found no evidence for a visual stimulation impact on serial biases, even under high statistical power conditions (n=112 subjects, >35000 trials). These contradictory findings suggest that different stimulation methods might in fact interfere with independent brain mechanisms and support the hypothesis of serial biases being produced by activity-silent mechanisms.

[30] This work was done in collaboration with Rebecca Martinez, which collected and helped analyze the TMS data under the supervision of Joao Barbosa, Josep Valls and Albert Compte.

Non-specific stimuli reactivate previous memories and increase serial biases in a computational model

This model was described in detail in the previous chapter, so we will only briefly summarize it here. We incorporated short-term plasticity in the bump-attractor model in order to simulate *activity-silent* mechanisms. To reactivate previous memories, we delivered an external input (*drive*) uniformly to all neurons (see S*imulating bump reactivation* in *Methods*) during the inter-trial period, when there was no stimulus representation in the neurons' firing rate but synapses were still facilitated as a result of the previous memory period persistent firing. To ensure such silent periods, we reset each ongoing simulation by delivering a strong negative input, effectively killing any remaining bump. After reset, delivering a *weak excitatory drive* resulted in the reactivation of previously active neurons by virtue of the selectivity still being imprinted in the facilitated synapses. On the other hand, delivering a *strong excitatory drive* saturated all the facilitated synapses, effectively removing any lingering synaptic tuning. As a behavioral proxy for memory reactivation, we applied our reactivation protocol during pre-stimulus and computed serial biases in the responses of the forthcoming trial. (see *Serial Biases* in *Methods*). This produced a non-linear dependence of the serial biases with drive strength (Figure 4.3.1a), which we tested experimentally using transmagnetic and visual stimulation (Figure 4.3.1b).

TMS-induced reactivations modulate serial biases

As a causal validation of fixation-period reactivation of dlPFC as a mechanism controlling serial biases, we designed a transcranial magnetic stimulation (TMS) perturbation study. This is a relevant experiment because the memory-dependent changes in EEG alpha-power distribution that we have analyzed so far (*Chapter 4.2*) depends to a great extent on parieto-occipital electrodes (Worden et al. 2000; Kelly et al. 2006; Medendorp et al. 2007; Foster et al. 2016; Foxe et al. 1998), which could pose a challenge to the correspondence of EEG findings with our monkey dlPFC data (Chapter 4.2). In non-invasive whole-brain measures, representations in larger and more organized occipital cortices might contribute more strongly to aggregate EEG measures but could be driven by top-down projections from association cortices (Reinhart et al. 2012), so we sought a specific proof of the causal involvement of dlPFC according to the predictions of our model and the monkey data (*Chapter 4.2*).

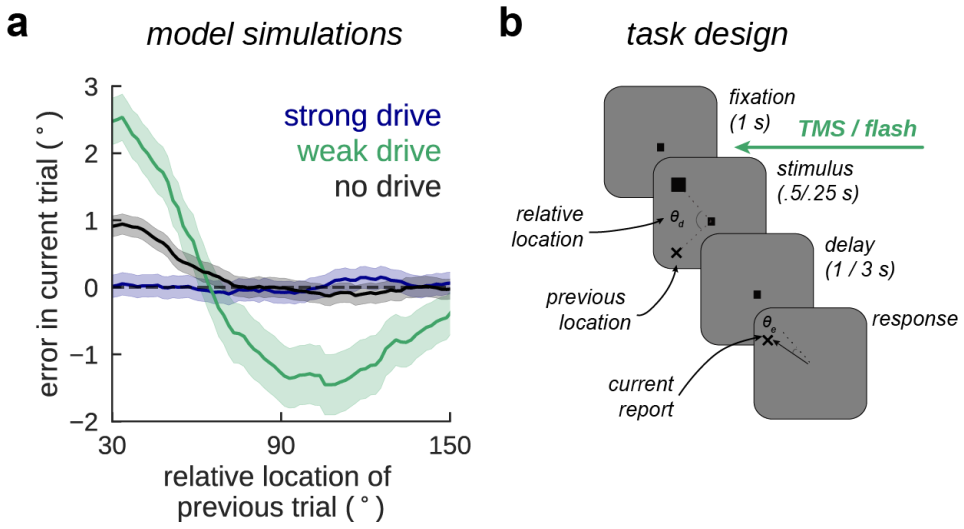


Figure 4.3.1. *Task design and predictions of the computational model.* **a)** We use a single-item visuo-spatial working memory task to test our predictions. On some trials, right before the stimulus presentation, we stimulated the subjects with either a targeted transmagnetic stimulation (TMS) or with a full-field flash. **b)** Behavioral responses computed from n=5000 simulations (for each curve) of the bump attractor model with short-term plasticity. A weak drive during pre-stimulus presentation increases serial biases, while a strong drive removes them.

Inspired by a previous study that reported reactivation of latent memories using a TMS protocol (Rose et al. 2016), we used TMS in 20 human subjects to causally link PFC to serial biases, and test the model prediction that reactivation-dependent behavioral biases would depend non-linearly on reactivation strength (Figure 4.3.1a). We sought to manipulate serial bias strength by stimulating PFC during the pre-cue period of our task using single-pulse TMS. We had two control conditions to test our hypotheses: (1) we targeted the TMS coil at dlPFC and vertex in interleaved blocks; and (2) we randomly chose TMS intensity, relative to the subject's resting motor threshold (RMT), in each trial (*sham:* 0%*, weak-tms:* 70%, and *strong-tms*: 130% of RMT, Methods). We found that TMS modulated serial biases when targeted at dlPFC but not at vertex (three-way interaction between sine of previous-current stimulus distance (*prev-curr*), TMS intensity and coil location, p=0.027. Two-way interaction between prev-curr and TMS intensity, for dlPFC p=0.034, for vertex p=0.97. *Methods, Regression models*). Importantly, our computational model prediction of a non-linear dependence with stimulation strength (Figure 4.3.1a) was supported by the data (ΔAIC=4.6, relative likelihood 0.9, for the comparison of regression models with non-linear vs. linear TMS-intensity factor. *Methods*).

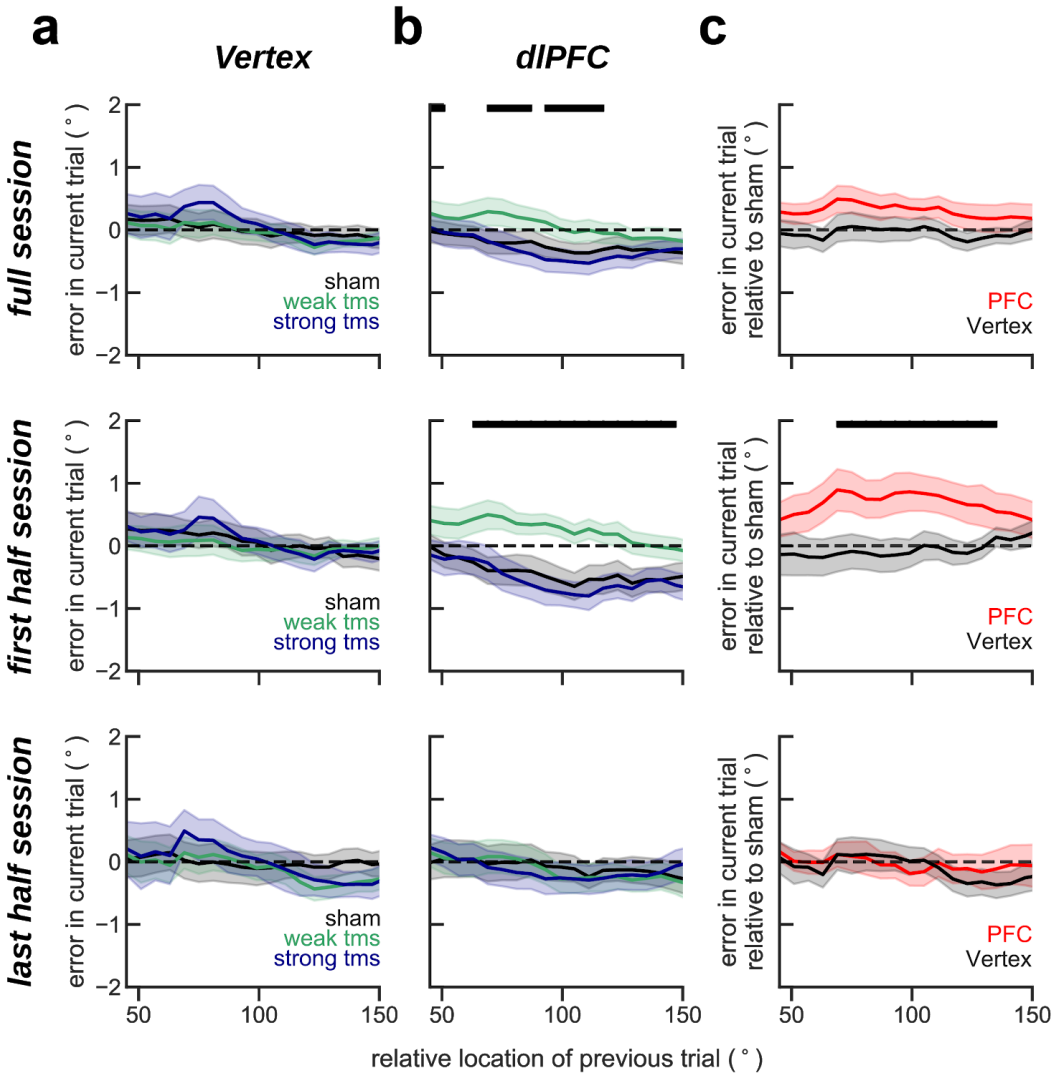


Figure 4.3.2. *Modulating serial biases with single-pulse TMS.* Serial biases computed using trials within **a)** vertex and **b)** PFC blocks, separating trials where the TMS pulse was strong (130% of resting motor threshold, blue) weak (70%, green) and sham (0%, black) during the first half each session. Serial biases were modulated by TMS in PFC, but not in vertex (p=0.011, three-way interaction with both blocks; p=0.0015, two-way interaction only for PFC blocks). **c)** Difference between sham and weak-tms trials for vertex and PFC blocks in black and red, respectively. Error-bars are bootstrapped standard errors of the mean; solid black lines on the top mark where the two curves are significantly different (one-sided permutation test at P=0.05).

Moreover, we found that the behavioral impact of TMS stimulation vanished through the session, as if subjects desensitized through repeated stimulation (Figure 4.3.2, Three-way interaction between prev-curr, TMS intensity and half-session p=0.02. Methods). Figure 4.3.2 shows serial biases for vertex (Figure 4.3.2a) and PFC blocks (Figure 4.3.2b) during the full session (Figure 4.3.2a, top) and first (Figure 4.3.2a,

middle) and last half (Figure 4.3.2a, bottom) of the experimental session (225 out of 450 trials per subject) Importantly, these are combined results from two separated experiments of n=10 subjects each, one being a registered replication (*Methods*. Figure 4.3.3). These results provide causal evidence for the involvement of PFC in the serial bias machinery during the pre-stimulus period. Further, we show that TMS impacts non-linearly serial biases, as predicted by our model simulations implementing the bump reactivation hypothesis via the interplay of bump attractor and activity-silent mechanisms.

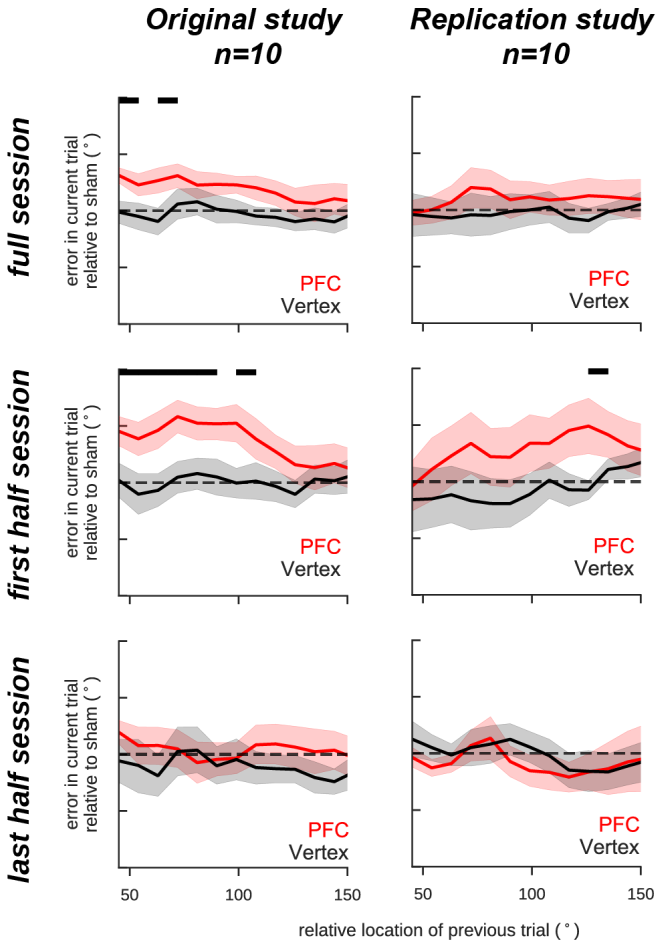


Figure 4.3.3. Same as Figure 4.3.2.c, but analysing data from original study (n=10) and replication study (n=10, https://osf.io/rguzn/) separately. Both studies have qualitatively similar effects, in particular for the first half where there was a TMS effect.

No evidence for visually-induced reactivations

In addition to transmagnetic stimulation, at least 2 studies report memory reactivation through a non-specific visual stimulus (Wolff et al. 2017; Wolff et al. 2015). We thus attempted to modulate serial biases using a full-field flash, by flickering the screen's background color between white and gray at two different frequencies (5 hz or 10 hz). Arguably, these two frequencies mirrored the two intensities predicted by our

model and the ones we used in the TMS experiments. We designed an experiment to be run in the online platform Amazon Mechanical Turks in order to obtain data from a large sample of participants and thus increase the statistics for what we expected to be a small effect. Figure 4.3.4a shows serial biases computed for subjects (n=112) with at least 100 correct trials. Computing serial biases separately by each flashing condition (no flash, 5 hz or 10 hz) did not reveal any significant serial bias modulation. If anything, a trend in the data suggested an unexpected reduction in serial biases with flashing stimulus, but this was not statistically significant and should be replicated in a new, larger study. Hence, in contrast to our model predictions and TMS experiments, our visual flash did not impact serial biases. This suggests that reactivation protocols using TMS and visual stimulus could be interfering with different brain mechanisms. In fact, TMS pulses can be targeted at specific brain areas, such as PFC here, but a visual stimulus has to travel through the whole brain hierarchy in order to produce such effect

.

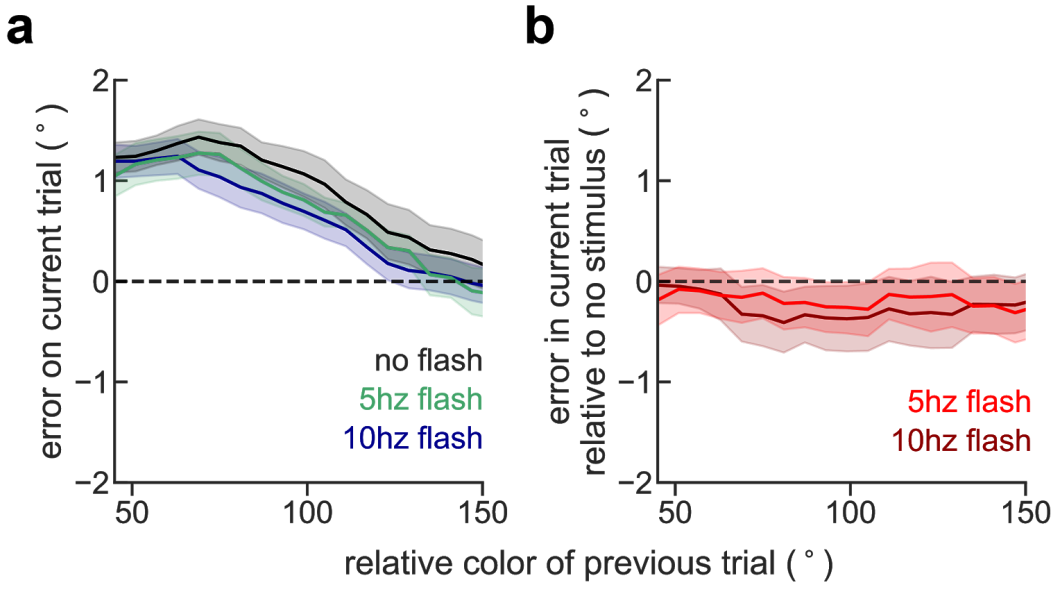


Figure 4.3.4. *No impact of a non-specific visual stimulation on serial biases.* **a)** Serial biases computed for subjects (n=112) with at least 100 correct trials (error < 45º). In contrast with the computer simulations and TMS experiments, there was no apparent effect of the non-specific stimulus on serial biases. **b)** Pairwise difference between flash (5hz and 10 hz) and no flash conditions shows a weak and non-significant (p>0.5, three-way interaction, See Methods, Linear models) negative modulation impact of the visual stimulus on serial biases.

Alternative explanations for “memory reactivation”[31]

Summary

Working memory maintenance is believed to be accomplished by reverberatory activity in high-order brain areas. Recently, an alternative hypothesis has gained substantial support. Under this hypothesis, neurons that were previously active during stimulus presentation can keep the stimulus information latent in their synapses, which connections were sculpted through fast, short-term plasticity. Theoretical work has shown that this latent code can be reactivated by means of a non-specific stimulus, inspiring the use of full-screen flashes or transcranial magnetic stimulation (TMS) in human neuroimaging experiments. Such studies, including our own (see 4.2), have interpreted an increase of stimulus information as a reactivation from a latent, activity-silent code. Critically, memory reactivation is by definition based on negative evidence: what could not be decoded then, can be decoded now. Here, we used computational simulations to argue for the existence of two alternative explanations for an increase in decodability following a non-specific stimulation. In light of our alternative hypotheses, previous negative results should be reinterpreted as an underlying weak code, undetectable using coarse neuroimaging signals. Under one alternative, a non-specific drive can strengthen the otherwise weak reverberatory activity. This code, too weak to be detected by typically coarse neuroimaging signals, can be further strengthened and become detectable when using neuroimaging techniques. Alternatively, a non-specific drive can decrease across-trial variability, which is in fact a strong, replicable phenomena that has been shown in many cortical areas of both monkeys and cats. This decrease in variance can trivially increase decodability without an actual signal increase (which is crucial in the ‘memory reactivation’ interpretation of the previous experimental results). Finally, we reanalyzed electroencephalograms (EEG) from 2 of those studies and found a potential confound for the one using a visual stimulus.

[31] This chapter contains parts of a larger, ongoing project in collaboration with Diego Lozano. Ideas laid down here are the result of long, fundamental discussions with Diego and Albert - therefore a collective effort. Nevertheless, included here are exclusively the simulations and analyses that were performed by me entirely.

A weak drive strengthens a weak bump

An increase in stimulus decodability after a non-specific stimulus can occur in networks that do not implement activity-silent mechanisms (e.g. short-term plasticity). An increase in stimulus tuning could happen because the non-specific drive pushes the network into a stable state with a higher maximal firing rate, without increasing the low firing rate state (illustration in Figure 4.3.5). To validate this intuition, we ran two sets of spiking-network simulations of the bump-attractor model without short-term plasticity (see *Bump-attractor model* in *Methods*). On some simulations, we drove the whole network uniformly during the last 0.5 s of the delay period. Figure 4.3.6a shows an example of such simulations, in which we applied a non-specific drive and Figure 4.3.6b another example simulation in which we did not apply the drive. As we expected, the persistent activity tuning for the driven network was higher than for the network without the drive (Figure 4.3.6c), despite the absence of activity-silent mechanisms. This effect becomes clearer when repeating the same analyses for 1000 network simulations (Figure 4.3.6d). We thus show that networks that do not directly implement activity-silent mechanisms can nevertheless show an increase of decodability when stimulated with a non-specific drive. This questions the interpretation that increases in stimulus decodability following unspecific stimulation implies the existence of activity-silent mechanisms (Wolff et al. 2017; Wolff et al. 2015; Rose et al. 2016).

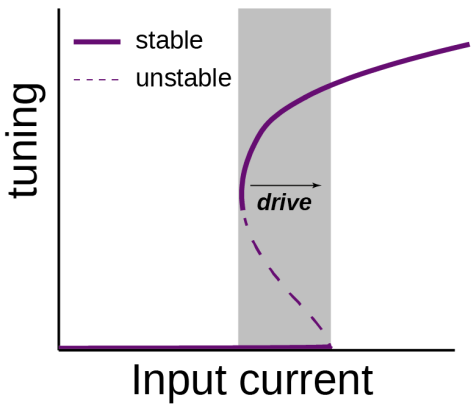


Figure 4.3.5. *Schematic illustration of how a weak drive can increase single trial tuning.* A weak bump (within the bistable gray area but close to the left bifurcation so it is sensitive to finite-size fluctuations) can be stabilized when pushed into a more stable state by a slight increase in input drive. Moreover, the firing rate tuning of a stable bump, well within the bistable gray region, can be further enhanced by a non-specific increase in input current without destroying the network bistability.

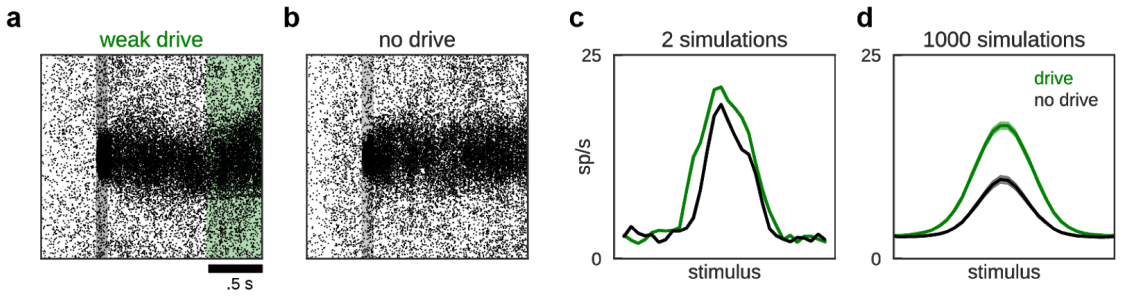


Figure 4.3.6. *A non-specific stimulus can increase tuning without reactivation.* Two example stimulations of a bump-attractor with **a)** and without **b)** a non-specific drive at the end of the trial. Importantly, we did not include short-term plasticity in either simulation. **c)** tuning during the last .5 s is higher for the trial in which a non-specific drive was delivered (green, a), compared to when no drive was delivered (black, b). **d)** Average tuning for 500 simulations with and without a drive delivered at the end of the delay.

A non-specific stimulus can increase signal-to-noise without reactivation

Increase of stimulus decodability following a non-specific input has been interpreted as evidence for an increase in stimulus information. However, decodability depends not only on the signal (μ), but also inversely on the amount of noise (σ). This relationship is commonly formalized as the signal-to-noise ratio: μ / σ. To illustrate this dependence, we ran 2 sets of simulations: 1) with stable noise (Figure 4.3.7, left) and 2) with a sudden drop in noise (Figure 4.3.7, right). Figure 4.3.7a shows 2 sets (left and right column, in Figure 4.3.7a) of 1000 simulations, each set representing the voltage recorded in one EEG channel during two stimulus conditions. The absolute difference between the signals during stimulus 1 and stimulus 2 should be interpreted as a decaying signal (μ) during the mnemonic period of a hypothetical working memory task. These are 2 versions of the same simulation, except for a decrease in variability (noise, σ) that we included on the right-hand side simulations (Figure 4.3.7b). Finally, Figure 4.3.7c shows how the *signal-to-noise* ratio can increase because of this noise decrease, rather than an increase in the signal that was the same for the 2 versions. These simple simulations thus show how an increase in decodability, previously seen in EEG experiments, can be trivially explained by a decrease in noise, instead of an increase in signal that is crucially invoked in the 'memory reactivation' interpretation of those experimental results (Wolff et al. 2017; Wolff et al. 2015; Rose et al. 2016).

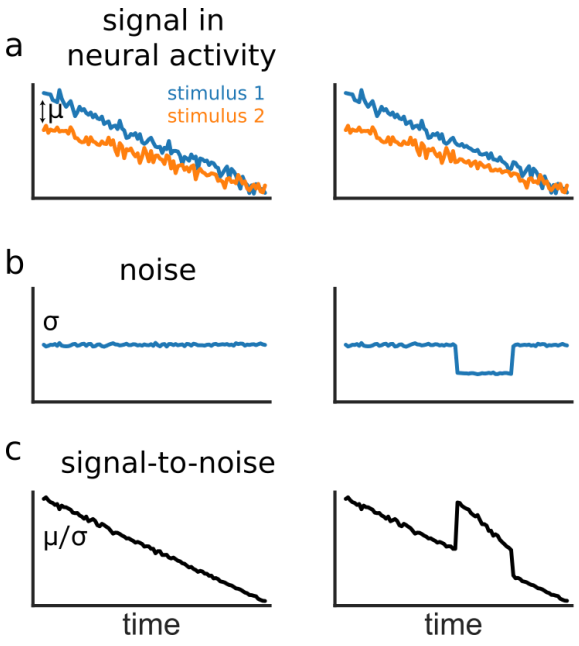


Figure 4.3.7. *A drop in variance induced by an external stimulus explains signal-to-noise increase.* **a)** 1000 trials without (right) and with (left) a drop in variance that could be induced by an external stimulus. Two stimuli conditions are simulated as two different slopes in the neural activity, effectively simulating a decaying code. **b)** Across-trial variance is stable (left) or drops temporarily (right). **c)** Decoding, measured as the signal-to-noise ratio, increases during the variance drop.

A non-specific visual stimulus decreases variance, while single-pulse TMS increases variance

In contrast with our first hypothesis, a signal-to-noise ratio (SNR) increase through a drop in variance can be easily tested in datasets of studies reporting evidence of memory reactivation in working memory tasks. We thus analysed across-trial variance in two EEG datasets. One in which the non-specific stimulus was a visual flash (Wolff et al. 2017) and another where it was a single transmagnetic stimulation (TMS) pulse (Rose et al. 2016). Figure 4.3.8 shows the across-trial variance, averaged after subtracting the across-trial variance computed .2 s before stimulus onset. In Fig 4.3.7a, there is a strong reduction (t-test, $p<0.005$) in variance for most subjects after the visual stimulus onset. Intriguingly, this reduction was not present in the dataset where TMS was used as a non-specific stimulus (Figure 4.3.8). Instead, we observed a strong increase in across-trial variance for all participants (t- test, $p<0.005$). These results discard our second hypothesis on the (Rose et al. 2016) dataset, where TMS was applied. However, a significant decrease in variability seen after a visual stimulus onset supports interpretations alternative to “memory reactivation” in experiments with visual pinging (Wolff et al. 2017; Wolff et al. 2015). Finally, we simulated 1000 trials of a bump-attractor network with short-term plasticity (as in *Chapter 4.2, Methods*). After storing one memory for 500ms, we reset the network activity with a strong, inhibitory input to all neurons. This effectively pushed

the activity back to baseline levels, but a synaptic-facilitation trace remained in the previously active neurons' synapses. Using a non-specific excitatory drive (as in *Chapter 4.2, Methods*), a bump was re-instantiated on *some* trials, leading to an increase of across-trial variability (Figure 4.3.8c) and decoding accuracy (as shown in Figure 4.2.5a).

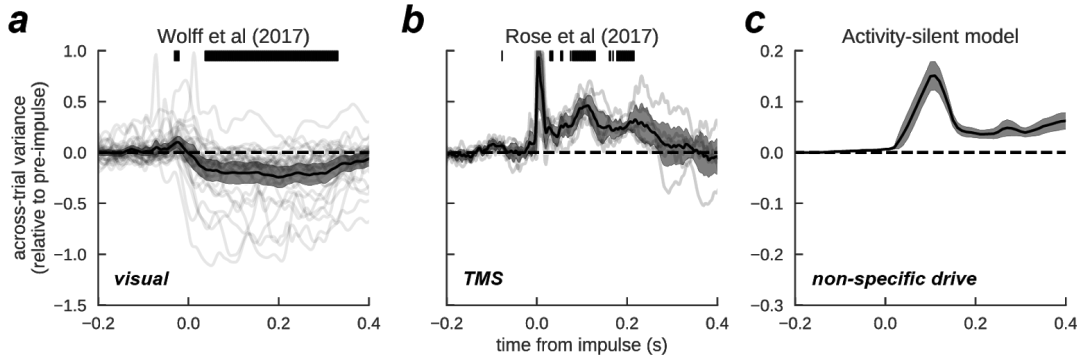


Figure 4.3.8. *Impulse-induced, across-trial variance change in human EEG.* Variance change relative to .2 s before impulse computed when the impulse was a visual stimulus **a)** and when it was a single-pulse TMS **b)**. Gray lines are changed in variance computed for single subjects. **c)** simulations of the bump-attractor model with short-term plasticity predict an increase of across-trial variance following a non-specific drive. Black solid bars mark where the change in variance was significant (t-test, $p<0.005$) and error-bars are bootstrapped 95% C.I. of the mean.

Interim conclusions

Inspired by previous studies that reported reactivation of latent memories using a non-specific stimuli, we used the same protocols to gather causal evidence for activity-silent mechanism as the source of serial biases. In particular, we attempted to reactivate previous memories either by using single-pulse TMS, which was targeted at dlPFC, or by presenting a strong, full-field visual flash. In both approaches, we stimulated in the pre-stimulus period. Unexpectedly, we got contradictory results. We were able to modulate serial biases when using single-pulse TMS (n=20 subjects), but failed when using a visual stimulus, even under high statistical power conditions (n=112 subjects, >35000 trials).

This motivated our final chapter, where we reinterpreted the aforementioned studies reporting short-term memory reactivation in the face of an increase in decoding accuracy. In this chapter, we provide two alternative explanations for the alleged memory reactivation findings. Crucially, neither alternatives require activity-silent mechanisms and propose instead that a weak, but active memory might be hidden in noisy neuroimaging or EEG signals. Using simple simulations we illustrate how a non-specific drive can strengthen an otherwise weak memory representation. Alternatively, a non-specific drive can decrease across-trial variability, which would trivially increase stimulus decodability. Finally, we reanalyzed the electroencephalograms (EEG) from those studies and compared them to simulations incorporating activity silent mechanisms. In line with our contradictory behavioral findings described above, only the TMS experiment was in accordance with such simulations.

5. DISCUSSION

Throughout this thesis, we investigated the neural circuit mechanisms of *visuo-spatial working memory* interference by formulating and sometimes testing predictions from a specific neural circuit hypothesis. In particular, our working hypothesis was the *bump-attractor* model, a neural network model which we updated along the way. First, by connecting two of such networks we overcome several limitations of the original model when used to simulate multi-item working memory tasks. Second, including short-term plasticity in the bump-attractor, we were able to explain neurophysiological findings previously interpreted as inconsistent with this model.
In sum, we reconcile two sources of disagreement: binding through a *temporal* (Singer 1999) vs *rate* code (Shadlen and Movshon 1999) and *activity-silent* (Stokes 2015) vs *activity-based* (Constantinidis et al. 2018; Barbosa 2017) working memory. On the one hand, our binding model is able to bind through synchronized oscillations (*temporal coding*), but encoding and decoding of associations is accomplished through *rate coding*. On the other hand, we incorporated activity-silent mechanisms in the bump attractor model, which expanded its explanatory power beyond mnemonic activity periods, including periods in which evidence for memory reactivations has been found.

Neural circuit basis of visuo-spatial working memory precision

In this chapter, our approach was to use a biologically constrained model to derive behavioral predictions. Specifically, we confirmed attractive and repulsive biases in the recollection of items located nearby in space. The attractive prediction is characteristic of other models of the same family of ours (Wei et al. 2012). That model, however, assumes global inhibition: regardless of its selectivity, an inhibitory neuron's impact onto another neuron remains the same. On the contrary, our model assumes local inhibition, that is the impact of one inhibitory neuron on another neuron decays with the selectivity difference between the two neurons (see Methods, Figure 3.3 global vs local inhibition). Importantly, only the model with local inhibition predicts repulsion for intermediate distances, which we validated using behavioral experiments. Our prediction and experimental result was more recently replicated in a larger sample size (Nassar et al. 2018; Chunharas et al. 2019), when memorizing colors and similar effects were also reported when remembering visual orientations (Rademaker et al. 2015). This supports our model as a general framework for

working memory. Conceptually, the very existence of interference effects has led some authors (Elmore et al. 2011; van den Berg et al. 2012) to interpret them as support for a resources model of working memory (Ma et al. 2014; Wilken and Ma 2004), which in its most basic formulation states that working memory can be seen as a resource shared between the memory representations of the different items. Indeed, similarity effects are not accommodated naturally in the alternative model, the slot model of working memory, which states that one memorizes each item independently until a maximal number of items is reached (Luck and Vogel 1997). As some authors have noted, however, similarity or interference effects would not pose any problem for the slots model if they primarily occurred in the encoding phase, not the mnemonic phase of the task (Johnson et al. 2009; Lin and Luck 2009). In our experiments, interference effects are not present when there is no delay period and the task is otherwise identical (Experiment 1 in the published study (Almeida et al. 2015), which is not shown in this thesis). This suggests that spatial interference of memorized locations occurs during the maintenance of information in working memory and not during the encoding of information, therefore supporting the resource model of working memory (see *Working memory capacity, Introduction* for a discussion of these two alternative models).

Finally, we do not claim that our model is able to simulate all task components. In particular, our task demanded the binding of two different features, color and location, while the model was only simulating the storage of position. This is justified by the lack of a consensual model for feature binding in working memory, but also because we demonstrated in behavioral experiments that the attraction and repulsion effects are independent of swap errors. However very rare in our experiment - perhaps due to low working memory load or iterm-item distances - swap errors have demonstrated to be ubiquitous in many other working memory tasks. Therefore, a complete understanding of this task requires explicitly simulating the binding component and, in fact, this was the motivation of our work discussed below.

Feature-binding in working memory through neuronal synchronization

Other models

During the development of our feature-binding model, at least two other types of neural models (Pina et al. 2018; Schneegans and Bays 2017a) were put forward - see also (Matthey et al. 2015) for a probabilistic account of (Schneegans and Bays

2017a). Both models from Paul Bays' laboratory (Matthey et al. 2015; Schneegans and Bays 2017a) assume that binding is accomplished through the so-called "conjunctive units" (see below), while the work of (Pina et al. 2018) implements binding by virtue of synchronized oscillatory activity. While similar in the approach, there are important differences between our model and the one by (Pina et al. 2018). In particular, they focused on the oscillatory behavior of discrete populations, while we modeled continuous attractors in spiking neural networks. First, because we simulate single spikes, our neuronal model is closer to biology, which we think is a relevant feature, in particular for our proof of concept claims (see below). Secondly, our continuous attractor model, rather than their discrete population model, keeps all the demonstrated explanatory power that is characteristic of these attractor models, such as explaining several behavioral biases seen in humans (Almeida et al. 2015; Nassar et al. 2018; Kilpatrick 2018; Kiyonaga et al. 2017) and monkeys (Papadimitriou et al. 2015; Funahashi et al. 1989); as well as explaining key neurophysiological dynamics during working memory maintenance periods, in humans (Edin et al. 2009; Kamiński et al. 2017) and monkeys (Wimmer et al. 2014; Sajad et al. 2016). This exploratory power was demonstrated once again with our simulations, as we explain several previous behavioral findings and derive a strong prediction from the central mechanism. Finally, modelling neuronal populations' firing rates as done in (Pina et al. 2018), instead of single neurons spike times as in our model, is a powerful approach to systematically explore the full parameter space and establish with great detail all the available regimes. On the other hand, our simulations show explicitly how a synchrony code can be controlled by firing rate inputs and can be read out from population firing rates, all without resorting to spike coincidence detectors; and it further shows the robustness of the mechanisms to noise inherent in spiking networks - both major concerns with binding-through-oscillation rate models such as (Pina et al. 2018) (see also problem 1 and 2, below).

Two opponent hypotheses

Moreover, (Pina et al. 2018) and (Schneegans and Bays 2017a) are good examples of a longstanding confrontation between two seemingly contradictory hypotheses regarding the underlying mechanism of feature-binding. The confrontation is as follows. From one perspective, binding can be accomplished simply by hard-wired anatomical connectivity, such that a neural population will naturally respond to a

bundle of features by receiving input from other, upstream populations, each of them selective to independent, still unbound features (Shadlen and Movshon 1999). This relates to how the brain encodes increasingly more complex objects, at least visual objects: one arbitrary area in the visual stream combines different features from upstream areas to generate a more complex feature. The application of this concept to the binding problem, however, has some limitations that remain unsolved. Namely, it implies that all the possible combinations of all the possible features have to be encoded *a priori* in hardwired connectivity. Is it reasonable to assume that there were hardwired neuronal structures representing all *The Garden of Earthly Delights* creatures in Hieronymus Bosch's brain? Is it reasonable to think that we need *those* hardwired structures to perceive his otherworldly creatures? It seems that under this theory we would need at least one cell in charge of each possible combination ("conjunctive units" as in (Schneegans and Bays 2017a)). This is of course a combinatorial problem that explodes quickly as we consider an increasing number of features. In the face of this limitation, proponents of the binding through synchronized activity argue that binding has to be supported mainly by neural *dynamics*, rather than neural *connectivity*. Crucially, neural dynamics can change on a much faster timescale than neural connectivity, an important requirement given our ability to quickly bind never-seen-together features, such as Bosch's mesmerizing creatures. On the other hand, opponents of the binding by synchrony (Shadlen and Movshon 1999) argue that this theory is biologically implausible because 1) such framework needs a *temporal encoder,* that tags bound features by a "temporal code" and a *temporal decoder*, that is able to distinguish which features are associated, both of which depending on undefined biological implementations; or as in (Shadlen and Movshon 1999) words: "each neuron therefore has to carry two distinct signals", one about the stimulus and other about the "tag". 2) clock-like synchronization, as intuitively required for this mechanism, is difficult to match with the noisy, typically asynchronous neuronal activity in the brain. As detailed below, our model solves problem 1) and suggests a solution to problem 2).

Strengths of our model

First (problem 1, above), in our model, only the maintenance of associations is accomplished through correlated oscillatory activity, in other words through a *temporal code*. Instead, encoding and decoding of associations is achieved through a *rate code*, by delivering flat pulses (i.e. without the need to be temporally precise) to

both the to-be-bound features exclusively (*encoding)* or just to one of them (*decoding*).

On the one hand, encoding the association between two different features through a pulse delivered simultaneously to each corresponding bump, resembles the sequential encoding hypothesis in working memory (Wolfe 1994; Bays, Gorgoraptis, et al. 2011). Moreover, there is evidence that a mechanism combining sequential and parallel encoding is implemented in the brain when solving multi-item working memory tasks (Bays, Gorgoraptis, et al. 2011). Our model implements such a combination. First, information about independent features arrives simultaneously to association areas from upstream sensory areas (note that we did not model sensory areas explicitly). Then, the correct associations are sequentially encoded by our simultaneous pulse, as it could be done by overtly attending to each stimulus sequentially (Schoenfeld et al. 2014). In fact, humans take longer to encode combined features than they take to encode the same amount of independent features (see *Binding of independent features* in *Introduction (1.2)*).

On the other hand, works modelling multi-item working memory though the storage of several bumps in a network (Krishnan et al. 2018; Wei et al. 2012; Nassar et al. 2018) - including our own (Almeida et al. 2015) - often used approaches that are biologically implausible to extract the location of one bump, while ignoring other simultaneously maintained bumps. Our approach, however, matches closely the "cueing" period of a multi-item working memory task, which consists of stimulating the "cued" locations while reading out from the whole color network population. The final behavioral output, for simplicity, is extracted by fitting a mixture of gaussians on the late-delay average activity of the color network and selecting the central value (color) of the gaussian component with higher amplitude. This algorithmic read-out could be replaced by a biologically plausible downstream network connected to the color circuit, and tuned to be in a winner-take-all regime - i.e. only able to maintain one bump at a time. However simple in essence, these encoding/decoding mechanisms overcome the first aforementioned limitation.

Second (problem 2, above), we found anti-phase dynamics within each network and phase-locking across networks, the central mechanisms for feature-binding in our model, to occur naturally in a broad range of parameters, indicating that the mechanisms proposed here are not the product of fine-tuning. Because our model is biologically constrained, it is a proof of concept that working memory binding through

synchronized activity is *at least* possible to occur in the brain. In fact, we simulated noisy integrate-and-fire neurons, supporting that the central mechanism implemented in our model has some degree of robustness to noise (but see below).

Limitations of our model

Finally, this study is limited in two ways. First, we did not simulate trials demanding binding of load 3. We expect that the main challenges associated with that improvement will be the encoding of more associations. Currently, we stimulated simultaneously only 1 pair of bumps - corresponding to only 1 association. In future work, we will study the conditions necessary for stimulating, sequentially, different bump pairs involved in all the associations (minimum of 3).

Another limitation is the oscillatory regime in which our model is operating, in which neurons are strongly synchronized with the population rhythm (Figure 4.1.8c). This regime, however derived from biologically constrained neuronal models, is arguably not biological itself. While there is abundant evidence that neuronal populations show strong oscillatory dynamics in working memory (e.g. (Pesaran et al. 2002)), single neuron dynamics approach a Poisson process (Softky and Koch 1993; Compte et al. 2003) - therefore not oscillatory at this scale. Early theoretical work (Brunel and Hakim 1999; Brunel and Wang 2003; Brunel 2000) has demonstrated that such oscillatory dynamics at the population level can coexist with noisy, unsynchronized neurons when randomly connected. It is however challenging to incorporate stable bump-attractors in such randomly connected networks (Hansel and Mato 2013), and for this reason we implemented all-to-all connectivity in our simulations. Future work should be done to study this limitation further, in particular by connecting randomly connected networks that do store multiple stable bump-attractors (Hansel and Mato 2013), but operating in anti-correlated oscillatory activity such as in our simulations.

Activity-silent mechanisms in PFC underlie serial biases in working memory

By studying the neural basis of serial biases in monkeys and humans, we have shown how the interplay of bump-activity dynamics and silent mechanisms in prefrontal cortex support behavioral biases in spatial working memory tasks. In these delayed-response tasks, prefrontal tuned persistent activity consistent with bump attractor dynamics characterizes the delay period and correlates with behavioral precision (Wimmer et al 2014). We have now seen that this sustained activation disappears from the prefrontal network between trials, and it reappears before the

new trial (Figures 4.2.1,4.2.4). We showed that this reactivation is causally implicated in the generation of behavioral serial biases (Figure 4.2.7 and Figure 4.3.2). Importantly, this reactivation is directly linked to activity recorded in the previous trial: it emerges specifically in those neural ensembles that show strongest persistent tuning in the delay (Figures 4.2.1c,d and 4.2.2), it is decoded from the human EEG with identical decoders (Figure 4.2.4), and it has the specific fingerprints of bump attractors as evaluated with pairwise spike-count correlations (Figure 4.2.3). These two disconnected periods of selective activations at the two ends of the inter-trial period are linked by activity-silent mechanisms in the prefrontal cortex, which carry selectivity from one trial to the next (Figure 4.2.6). Taken together, our results are consistent with the view that attractor-based and activity-silent mechanisms are jointly represented in the local prefrontal circuit and their tight interplay conjointly supports behavior in spatial working memory (Fujisawa et al. 2008). We specified this view in a computational network model: delay-period attractor dynamics load selective activity-silent mechanisms, which then retain information between trials and allow reactivations into recapitulating attractors (Figure 4.2.5).

Our data provides experimental support in prefrontal circuits that non-specific stimulation can retrieve information maintained subthreshold, similar to the modeling ideas put forward by (Mongillo et al. 2008) This goes beyond previous studies using neuroimaging and EEG methods (Rose et al. 2016; Wolff et al. 2017; Wolff et al. 2015), because by interrogating single-neuron data we can truly validate that information is absent from firing rates but still present in synchrony parameters, thus revealing a latent subthreshold tuning, and this tuning can be reinstated in firing rates when triggered by external events. This crucial test of memory reactivation from activity-silent sources was still lacking experimentally. However, our data also supports the idea that activity-silent and attractor-based mechanisms are not orthogonal, alternative mechanisms but they are largely co-expressed in the circuit and underlie different behaviors: while persistent attractor-based activity is engaged during active maintenance of memories, activity-silent maintenance supports instead secondary, possibly involuntary memory traces, here expressed in small serial biases. Similar ideas have been proposed in the context of prioritized and unprioritized memories (Rose et al. 2016; Wolff et al. 2017). Note, however, that in our proposed framework the close interplay between attractor-based and activity-silent mechanisms does not allow to protect activity-silent memories from

intervening attractor-based activations in the circuit. This would therefore predict that in the prioritization protocols of (Rose et al. 2016) specific patterns of interferences in latent non-prioritized memories should be observed, depending on the duration and distance of simultaneous prioritized memories.

An important point in our study is that the proposed mechanisms can be directly linked with behavioral parameters in our task. Indeed, we found robust evidence for the role of bump reactivations from activity-silent traces in generating working memory serial biases. This is a significant advancement over previous studies (Sugase-Miyamoto et al. 2008; Fujisawa et al. 2008) that explicitly demonstrates the possible behavioral impact of activity-silent traces.

We propose a computational model that can parsimoniously explain our data using short-term facilitation in the synapses of a recurrent network. Our findings however are not univocally identifying this mechanism and we could have chosen another sub-threshold mechanism with a long time constant to implement our hypothesis computationally (e.g. calcium-activated depolarizing currents (Tegnér et al. 2002), depolarization-induced suppression of inhibition (Carter and Wang 2007). Still, several lines of evidence support the involvement of short-term plasticity in prefrontal function: there is evidence for enhanced short-term facilitation among PFC neurons (Wang et al. 2006), and neural activity in rodent medial PFC presents functional connectivity patterns highly consistent with short-term plasticity dynamics (Fujisawa et al. 2008). We also note that short-term plasticity has also been used in previous computational models of interacting activity-based and activity-silent dynamics (Mongillo et al. 2008) and of serial biases (Kilpatrick 2018; Bliss and D'Esposito 2017).

Build-up of serial biases in color working memory

We provide the first evidence of serial dependence in color working memory. Serial dependence had been characterized with great detail for other visual features (Kiyonaga et al. 2017), and in particular in spatial working memory (Bliss et al. 2017; Papadimitriou et al. 2015; Papadimitriou et al. 2017). Several common features of color and spatial working memory suggest that serial dependence could also be similar in color: simultaneously memorized stimuli interfere attractively when presented at close distances (Almeida et al. 2015; Nassar et al. 2018), and memory

precision decreases with memory period duration (Zhang and Luck 2009; Nilsson and Nelson 1981; Bliss et al. 2017). These commonalities are in contrast with the differences of neural representations. While spatial representations consolidate early in the visual pathway (Wandell et al. 2007), complex transformations in color representations occur as color information travels from the photoreceptors in the retina, to visual cortex, and into association cortex (Johnson and Mullen 2016). The fact that serial dependence is similar for color and spatial working memory thus suggests that it depends on inter-trial interferences that occur at processing stages with representational maps equally distant from the corresponding perceptual map, and this points at higher color processing stages. A candidate region for this is the inferotemporal (IT) cortex, where continuous neuronal representations of color of circular shape on the two perceptual cardinal axes (yellowish-bluish and greenish-reddish axis) have been found (Bohon et al. 2016; Chang et al. 2017).

The analogy of color and angular location neural representations motivated us to simulate color working memory similarly to spatial working memory of angular locations (Compte et al. 2000). We simulated the angular memory trace in the memory period as a diffusion process (Wimmer et al. 2014) with a drift toward the previous trial memory trace that introduces serial dependence (Kilpatrick 2018). We used this model test the concerns about the impact of systematic biases in the estimation of serial dependence. This is a general concern that has been raised for other visual features (Samaha et al. 2018; Bliss et al. 2017; Papadimitriou et al. 2017; Papadimitriou et al. 2015), but in the case of color it may be particularly important for the marked perceptual systematic biases that have been reported (Bae et al. 2014). We therefore incorporated strong systematic biases in the reports of our model simulations, we tested the impact on the estimation of serial dependence and we developed new analysis strategies to address this. One typical strategy for systematic bias removal is to low-pass filter the responses as a function of stimulus feature (Bliss et al. 2017; Papadimitriou et al. 2017; Papadimitriou et al. 2015; Samaha et al. 2018). This approach depends on parameters that are often subjectively decided (e.g. size of sliding window). In addition, removing systematic biases incorrectly, for example when subjects do not have systematic biases, can introduce extra biases in otherwise clean data. We showed that by folding the serial dependence function, one can reduce the impact of systematic biases on serial dependence without adding biases in unbiased data and without specifying arbitrary

parameter values. We therefore conclude that this analysis allows a more robust estimation of serial dependence in behavioral studies.

Theoretical models have proposed that short-term subthreshold mechanisms in inter-trial intervals underlie serial dependence in delayed-estimation of location (Carter and Wang 2007; Kilpatrick 2018; Bliss and D'Esposito 2017). In this class of models, neural activity in previous trial's mnemonic representations engage plasticity mechanisms that leave a selective trace in the network's synapses. This trace interferes with neural activations in the next trial by biasing the neural representation of the new stimulus towards the previous memorized location. These dynamics explain most experimental findings of serial dependence in spatial working memory (Kilpatrick 2018), and our simplified modeling approach is consistent with this mechanistic substrate (Kilpatrick 2018). However, our finding that attractive serial biases build up in the course of an experimental session is not explained by these models. Indeed, the short-term synaptic plasticity mechanisms invoked so far operate in time scales of a few seconds, much shorter than the time scale of the experimental session. Our results reveal that additional mechanisms, accumulating in a time scale of 10's of minutes or hours, are also responsible for the instantiation of serial dependencies in delayed-estimation of color tasks. Possible mechanisms are changes in plasticity efficacy itself (i.e. "metaplasticity"), modulating synaptic release probability over the experimental session. We tentatively speculate that habit-related endocannabinoid modulation of synaptic release (Castillo et al. 2012; Carter and Wang 2007) could mediate serial dependence build-up as a non-adaptive result of task habituation.

The build-up of serial dependencies during an experimental session has further implications for how we interpret their functional role. If serial dependence was an adaptation to exploit the world's tendency to remain stable (Cicchini et al. 2018), and this adaptation could occur in the time scale of hour fractions (as recently shown for systematic biases in delayed-estimation of color tasks (Panichello et al. 2018)), memorizing a sequence of uncorrelated stimuli should decrease serial dependence in the course of an experimental session. Alternatively, if hard-wired mechanisms underlie serial dependence, we wouldn't expect any change. Instead, our results show that serial dependence builds up, suggesting that it does not respond to an active adaptation to the statistics of visual stimuli in the environment (at least in the

time scale of hours) but instead may reflect a plasticity mechanism driven by repeated selective neuronal activations in the circuit.

Reactivation of previous-trial memories with non-specific stimuli

In this study, we tested two different, but related hypotheses: 1) activity-silent mechanisms during the inter-trial period underlie serial biases and 2) non-specific stimuli are effective methods of memory reactivation. We validated these hypotheses using TMS, but failed when using non-specific visual stimuli, even under high statistical power conditions (n=112 subjects, >35000 trials). Moreover, TMS had a strong impact on behavior that matched tightly our neural model simulations: weak stimulation increased serial biases, while strong stimulation removed serial biases. Critically, this effect was observed in blocks where dlPFC, but not vertex, was stimulated. Intriguingly, within blocks where the TMS coil was targeted at dlPFC and considering only trials where no TMS was applied (i.e. within dlPFC sham condition), behavioral responses were strongly repelled from the location in the previous trial. This did not happen in Vertex, where serial biases resembled the profiles that have been typically reported in the literature, with significant attractive serial biases. We speculate that this was due to carry-over effects from previous TMS-stimulated trials. In fact, studies combining TMS with single unit records report that fast, excitatory TMS effect can be followed by a slow, inhibitory effect (Romero et al. 2018; Murphy et al. 2016). Future work involving more TMS intensities and carefully controlled block designs will be necessary to clarify these results further. Nevertheless, because this behavioral effect was exclusive of dlPFC, our results undoubtedly show that dlPFC hosts a key component of the serial bias machinery. On the other hand, our negative results when using visual stimuli, instead of TMS, suggests that these 2 techniques that provided similar decoding results might, in fact, interfere with two different brain mechanisms - a hypothesis that we endorse in the final project of this thesis.

Alternative explanations for "memory reactivation"

Increase in decodability following a non-specific stimulus has been regarded as evidence for memory reactivation (Wolff et al. 2017; Wolff et al. 2015; Rose et al. 2016), especially when this increase departs from change levels (Rose et al. 2016). Here, we introduced two very simple alternative interpretations of those previous findings. We did that by using simple proof-of-concept simulations that suggest at

least two alternative scenarios, in which memory reactivation is not a necessary condition. In particular, we showed how a non-specific stimulus can increase a single neuron's tuning without any activity-silent mechanism implemented on their synapses. In our simulations, in contrast with aforementioned neuroimaging studies, decodability is weak, but never absent. We assume that stimulus representation during the delay is too weak to be detected by neuroimaging techniques, such as EEG and fmri (Wolff et al. 2017; Wolff et al. 2015; Rose et al. 2016). This is not an unreasonable assumption, since these negative results come exclusively from neuroimaging studies, but stimulus-selective delay-activity has been replicated in many studies based on single unit activity (Christophel et al. 2017; Leavitt et al. 2017), even when using similar tasks (Barbosa 2017; Spaak et al. 2017; Watanabe and Funahashi 2014). In fact, (Christophel et al. 2018) showed that a lack of decodability from previous neuroimaging studies (Wolff et al. 2017; Wolff et al. 2015; Rose et al. 2016) can be overcome by a substantial increase of power (n=87 subjects and more than 16000 trials). Alternatively, we argued that an increase in decodability could be a spurious consequence of a decrease in across-trial variability. In fact, we analysed electroencephalograms from the 2 studies in question and we found a decrease in across-trial variability in the study that used a visual, non-specific stimuli, but not for the one using transmagnetic stimulation. This contradiction might explain why we could modulate serial biases stimulating PFC with single-pulse TMS, but failed when using a full-field flash. Future work involving simultaneous EEG and non-specific stimulation, both with TMS and visual flash, is necessary to confirm this hypothesis.

6. CONCLUSIONS

Chapter 4.1: Interference from simultaneous memories

1. We validated two behavioral predictions of the bump-attractor model, when storing multiple items. In particular, we found that humans have repulsive and attractive biases, as predicted by the model.

2. Furthermore, aiming to account for swap-errors, other sources of biases in multi-item working memory experiments, we extended the classical bump-attractor model. Our biologically-constrained model offers a plausible mechanism for feature-binding through selective synchronization. Importantly, it explains when this feature binding fails, including how it depends on delay duration and inter-item distances. Moreover, it provides a strong, testable prediction from its central underlying mechanism - phase-locked oscillatory activity during the memory periods.

Chapter 4.2: Interference from previous memories

3. Combining humans and monkey neurophysiological experiments, we found that after memory-guided reports, past stimulus information ceased to be represented in PFC neurons. Surprisingly, this information was again represented by such neurons before furthcoming stimulus onset. Because this was reminiscent of 'activity-silent' mechanisms, we called this phenomenon *reactivation*. By introducing short-term plasticity dynamics in the bump-attractor model, we explained such reactivation dynamics and derived novel predictions that we then validated, linking behavior and neurophysiology. First, we found enhanced functional connectivity between PFC simultaneously recorded neurons that were involved in memory maintenanced, compared with trials that these same neurons were not involved in memory maintenanced. Second, in line with the simulations, we found that during trials undergoing higher information reactivation, monkeys and humans were prone to stronger serial biases. Together, these results point to an activity-silent mechanism, such as short-term plasticity, as the underlying mechanism of interference from previous memories in working memory (i.e. serial biases).

4. Finally, we described for the first time serial biases in color working memory. In particular, we found that attractive serial biases build up in the course of an experimental session, an effect that cannot be explained solely by short-term plasticity. Rather, to account for such effect, we tentatively speculate that metaplasticity mechanisms need to be included in existing theoretical model for working memory interference.

Chapter 4.3: Reactivation of previous memories with non-specific stimuli

5. Inspired by previous studies that reported reactivation of latent memories using a non-specific stimuli, we used the same protocols to gather causal evidence for activity-silent mechanism as the source of serial biases. In particular, we attempted to reactivate previous memories either by using single-pulse TMS, which was targeted at dlPFC, or by presenting a strong, full-field visual flash. In both approaches, we stimulated during the pre-stimulus period. Unexpectedly, we got contradictory results. We were able to modulate serial biases when using single-pulse TMS (n=20 subjects), but failed when using a visual stimulus, even under high statistical power conditions (n=112 subjects, >35000 trials).

6. This motivated our final chapter, where we reinterpreted the aforementioned studies reporting short-term memory reactivation in the face of an increase in decoding accuracy. In this chapter, we provide two alternative explanations for the alleged memory reactivation findings. Crucially, neither alternatives require activity-silent mechanisms and propose instead that a weak, but active memory might be hidden in noisy neuroimaging or EEG signals. Using simple simulations we illustrate how a non-specific drive can strengthen an otherwise weak memory representation. Alternatively, a non-specific drive can decrease across-trial variability, which would trivially increase stimulus decodability. Finally, we reanalyzed the electroencephalograms (EEG) from those studies and compared them to simulations incorporating activity silent mechanisms. In line with our contradictory behavioral findings described above, only the TMS experiment was in accordance with such simulations.

Bibliography

Abbott, L.F. and Regehr, W.G. 2004. Synaptic computation. *Nature* 431(7010), pp. 796–803.

Abbott, L.F., Varela, J.A., Sen, K. and Nelson, S.B. 1997. Synaptic depression and cortical gain control. *Science* 275(5297), pp. 220–224.

Adam, K.C.S., Vogel, E.K. and Awh, E. 2017. Clear evidence for item limits in visual working memory. *Cognitive Psychology* 97, pp. 79–97.

Aertsen, A.M., Gerstein, G.L., Habib, M.K. and Palm, G. 1989. Dynamics of neuronal firing correlation: modulation of “effective connectivity”. *Journal of Neurophysiology* 61(5), pp. 900–917.

Alais, D., Kong, G., Palmer, C. and Clifford, C. 2018. Eye gaze direction shows a positive serial dependency. *Journal of Vision* 18(4), p. 11.

Alexi, J., Cleary, D., Dommisse, K., Palermo, R., Kloth, N., Burr, D. and Bell, J. 2018. Past visual experiences weigh in on body size estimation. *Scientific reports* 8(1), p. 215.

Almeida, R., Barbosa, J. and Compte, A. 2015. Neural circuit basis of visuo-spatial working memory precision: a computational and behavioral study. *Journal of Neurophysiology* 114(3), pp. 1806–1818.

Amarasingham, A., Harrison, M.T., Hatsopoulos, N.G. and Geman, S. 2012. Conditional modeling and the jitter method of spike resampling. *Journal of Neurophysiology* 107(2), pp. 517–531.

Bae, G.-Y., Olkkonen, M., Allred, S.R., Wilson, C. and Flombaum, J.I. 2014. Stimulus-specific variability in color working memory with delayed estimation. *Journal of Vision* 14(4).

Barak, O. and Tsodyks, M. 2007. Persistent activity in neural networks with dynamic synapses. *PLoS Computational Biology* 3(2), p. e35.

Barak, O., Tsodyks, M. and Romo, R. 2010. Neuronal population coding of parametric working memory. *The Journal of Neuroscience* 30(28), pp. 9424–9430.

Barbosa, J. 2017. Working memories are maintained in a stable code. *The Journal of Neuroscience* 37(35), pp. 8309–8311.

Barrouillet, P., De Paepe, A. and Langerock, N. 2012. Time causes forgetting from working memory. *Psychonomic Bulletin & Review* 19(1), pp. 87–92.

Bates, D., Mächler, M., Bolker, B. and Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software* 67(1), pp. 1–48.

Bays, P.M. 2016. Evaluating and excluding swap errors in analogue tests of working memory. *Scientific reports* 6, p. 19203.

Bays, P.M., Catalao, R.F.G. and Husain, M. 2009. The precision of visual working

memory is set by allocation of a shared resource. *Journal of Vision* 9(10), p. 7.1-11.

Bays, P.M., Gorgoraptis, N., Wee, N., Marshall, L. and Husain, M. 2011. Temporal dynamics of encoding, storage, and reallocation of visual working memory. *Journal of Vision* 11(10).

Bays, P.M., Wu, E.Y. and Husain, M. 2011. Storage and binding of object features in visual working memory. *Neuropsychologia* 49(6), pp. 1622–1631.

van den Berg, R., Shin, H., Chou, W.-C., George, R. and Ma, W.J. 2012. Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America* 109(22), pp. 8780–8785.

Bettencourt, K.C. and Xu, Y. 2016. Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nature Neuroscience* 19(1), pp. 150–157.

Blake, R., Cepeda, N.J. and Hiris, E. 1997. Memory for visual motion. *Journal of Experimental Psychology. Human Perception and Performance* 23(2), pp. 353–369.

Bliss, D.P. and D'Esposito, M. 2017. Synaptic augmentation in a cortical circuit model reproduces serial dependence in visual working memory. *Plos One* 12(12), p. e0188927.

Bliss, D.P., Sun, J.J. and D'Esposito, M. 2017. Serial dependence is absent at the time of perception but increases in visual working memory. *Scientific reports* 7(1), p. 14739.

Bohon, K.S., Hermann, K.L., Hansen, T. and Conway, B.R. 2016. Representation of perceptual color space in macaque posterior inferior temporal cortex (the V4 complex). *eNeuro* 3(4).

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S. and Movshon, J.A. 1996. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience* 13(1), pp. 87–100.

Brouwer, G.J. and Heeger, D.J. 2009. Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience* 29(44), pp. 13992–14003.

Brunel, N. 2000. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *Journal of Computational Neuroscience* 8(3), pp. 183–208.

Brunel, N. and Hakim, V. 1999. Fast global oscillations in networks of integrate-and-fire neurons with low firing rates. *Neural Computation* 11(7), pp. 1621–1671.

Brunel, N. and van Rossum, M.C.W. 2007. Lapicque's 1907 paper: from frogs to integrate-and-fire. *Biological Cybernetics* 97(5–6), pp. 337–339.

Brunel, N. and Wang, X.-J. 2003. What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and

excitation-inhibition balance. *Journal of Neurophysiology* 90(1), pp. 415–430.

Burnham, K.P. and Anderson, D.R. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological methods & research* 33(2), pp. 261–304.

Carter, E. and Wang, X.-J. 2007. Cannabinoid-mediated disinhibition and working memory: dynamical interplay of multiple feedback mechanisms in a continuous attractor model of prefrontal cortex. *Cerebral Cortex* 17 Suppl 1, pp. i16-26.

Castillo, P.E., Younts, T.J., Chávez, A.E. and Hashimotodani, Y. 2012. Endocannabinoid signaling and synaptic function. *Neuron* 76(1), pp. 70–81.

Cecchi, G.A., Rao, A.R., Xiao, Y. and Kaplan, E. 2010. Statistics of natural scenes and cortical color processing. *Journal of Vision* 10(11), p. 21.

Chang, L., Bao, P. and Tsao, D.Y. 2017. The representation of colored objects in macaque color patches. *Nature Communications* 8(1), p. 2064.

Chang, M.H., Armstrong, K.M. and Moore, T. 2012. Dissociation of response variability from firing rate effects in frontal eye field neurons during visual stimulation, working memory, and attention. *The Journal of Neuroscience* 32(6), pp. 2204–2216.

Chelazzi, L., Miller, E.K., Duncan, J. and Desimone, R. 2001. Responses of neurons in macaque area V4 during memory-guided visual search. *Cerebral Cortex* 11(8), pp. 761–772.

Christophel, T.B., Iamshchinina, P., Yan, C., Allefeld, C. and Haynes, J.-D. 2018. Cortical specialization for attended versus unattended working memory. *Nature Neuroscience* 21(4), pp. 494–496.

Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R. and Haynes, J.-D. 2017. The distributed nature of working memory. *Trends in Cognitive Sciences* 21(2), pp. 111–124.

Chunharas, C., Rademaker, R.L., Brady, T.F. and Serences, J. 2019. Adaptive memory distortion in visual working memory.

Cicchini, G.M., Anobile, G. and Burr, D.C. 2014. Compressive mapping of number to space reflects dynamic encoding mechanisms, not static logarithmic transform. *Proceedings of the National Academy of Sciences of the United States of America* 111(21), pp. 7867–7872.

Cicchini, G.M., Mikellidou, K. and Burr, D. 2017. Serial dependencies act directly on perception. *Journal of Vision* 17(14), p. 6.

Cicchini, G.M., Mikellidou, K. and Burr, D.C. 2018. The functional role of serial dependence. *Proceedings. Biological Sciences / the Royal Society* 285(1890).

Cisek, P. 2006. Integrated neural processes for defining potential actions and deciding between them: a computational model. *The Journal of Neuroscience* 26(38), pp. 9761–9770.

Cisek, P. and Kalaska, J.F. 2005. Neural correlates of reaching decisions in dorsal premotor cortex: specification of multiple direction choices and final selection of action. *Neuron* 45(5), pp. 801–814.

Compte, A., Brunel, N., Goldman-Rakic, P.S. and Wang, X.J. 2000. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cerebral Cortex* 10(9), pp. 910–923.

Compte, A., Constantinidis, C., Tegner, J., Raghavachari, S., Chafee, M.V., Goldman-Rakic, P.S. and Wang, X.-J. 2003. Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *Journal of Neurophysiology* 90(5), pp. 3441–3454.

Constantinidis, C., Franowicz, M.N. and Goldman-Rakic, P.S. 2001a. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex. *The Journal of Neuroscience* 21(10), pp. 3646–3655.

Constantinidis, C., Franowicz, M.N. and Goldman-Rakic, P.S. 2001b. The sensory nature of mnemonic representation in the primate prefrontal cortex. *Nature Neuroscience* 4(3), pp. 311–316.

Constantinidis, C., Funahashi, S., Lee, D., Murray, J.D., Qi, X.-L., Wang, M. and Arnsten, A.F.T. 2018. Persistent spiking activity underlies working memory. *The Journal of Neuroscience* 38(32), pp. 7020–7028.

Constantinidis, C. and Goldman-Rakic, P.S. 2002. Correlated discharges among putative pyramidal neurons and interneurons in the primate prefrontal cortex. *Journal of Neurophysiology* 88(6), pp. 3487–3497.

Constantinidis, C., Williams, G.V. and Goldman-Rakic, P.S. 2002. A role for inhibition in shaping the temporal flow of information in prefrontal cortex. *Nature Neuroscience* 5(2), pp. 175–180.

Cowan, N. 2001. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24(1), p. 87–114; discussion 114.

Czoschke, S., Fischer, C., Beitner, J., Kaiser, J. and Bledowski, C. 2018. Two types of serial dependence in visual working memory. *British Journal of Psychology*.

Delvenne, J. and Bruyer, R. 2004. Does visual short‐term memory store bound features? *Visual cognition* 11(1), pp. 1–27.

Dingledine, R., Borges, K., Bowie, D. and Traynelis, S.F. 1999. The glutamate receptor ion channels. *Pharmacological Reviews* 51(1), pp. 7–61.

Dittman, J.S., Kreitzer, A.C. and Regehr, W.G. 2000. Interplay between facilitation, depression, and residual calcium at three presynaptic terminals. *The Journal of Neuroscience* 20(4), pp. 1374–1385.

Druckmann, S. and Chklovskii, D.B. 2012. Neuronal circuits underlying persistent representations despite time varying activity. *Current Biology* 22(22), pp. 2095–2103.

Dubois, J., de Berker, A.O. and Tsao, D.Y. 2015. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. *The Journal of Neuroscience* 35(6), pp. 2791–2802.

Durstewitz, D., Seamans, J.K. and Sejnowski, T.J. 2000. Neurocomputational models

of working memory. *Nature Neuroscience* 3 Suppl, pp. 1184–1191.

Edin, F., Klingberg, T., Johansson, P., McNab, F., Tegnér, J. and Compte, A. 2009. Mechanism for top-down control of working memory capacity. *Proceedings of the National Academy of Sciences of the United States of America* 106(16), pp. 6802–6807.

Elmore, L.C., Ma, W.J., Magnotti, J.F., Leising, K.J., Passaro, A.D., Katz, J.S. and Wright, A.A. 2011. Visual short-term memory compared in rhesus monkeys and humans. *Current Biology* 21(11), pp. 975–979.

Emrich, S.M. and Ferber, S. 2012. Competition increases binding errors in visual working memory. *Journal of Vision* 12(4).

Fischer, J. and Whitney, D. 2014. Serial dependence in visual perception. *Nature Neuroscience* 17(5), pp. 738–743.

Foster, J.J., Bsales, E.M., Jaffe, R.J. and Awh, E. 2017. Alpha-Band Activity Reveals Spontaneous Representations of Spatial Position in Visual Working Memory. *Current Biology* 27(20), p. 3216–3223.e6.

Foster, J.J., Sutterer, D.W., Serences, J.T., Vogel, E.K. and Awh, E. 2016. The topography of alpha-band activity tracks the content of spatial working memory. *Journal of Neurophysiology* 115(1), pp. 168–177.

Fougnie, D. and Alvarez, G.A. 2011. Object features fail independently in visual working memory: evidence for a probabilistic feature-store model. *Journal of Vision* 11(12).

Foxe, J.J., Simpson, G.V. and Ahlfors, S.P. 1998. Parieto-occipital approximately 10 Hz activity reflects anticipatory state of visual attention mechanisms. *Neuroreport* 9(17), pp. 3929–3933.

Fritsche, M., Mostert, P. and de Lange, F.P. 2017. Opposite effects of recent history on perception and decision. *Current Biology* 27(4), pp. 590–595.

Fujisawa, S., Amarasingham, A., Harrison, M.T. and Buzsáki, G. 2008. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nature Neuroscience* 11(7), pp. 823–833.

Funahashi, S., Bruce, C.J. and Goldman-Rakic, P.S. 1989. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology* 61(2), pp. 331–349.

Fuster, J.M. and Alexander, G.E. 1971. Neuron activity related to short-term memory. *Science* 173(3997), pp. 652–654.

Gayet, S., Paffen, C.L.E. and Van der Stigchel, S. 2018. Visual working memory storage recruits sensory processing areas. *Trends in Cognitive Sciences* 22(3), pp. 189–190.

Gerstner, W., Kistler, W.M., Naud, R. and Paninski, L. 2014. *Neuronal dynamics: from single neurons to networks and models of cognition*. Cambridge: Cambridge University Press.

Gold, J.I., Law, C.-T., Connolly, P. and Bennur, S. 2008. The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *Journal of Neurophysiology* 100(5), pp. 2653–2668.

Goldman, M.S. 2009. Memory without Feedback in a Neural Network. *Neuron* 93(3), p. 715.

Goldman-Rakic, P.S. 1993. Working memory and the mind. . *Scientific American*, pp. 67–77.

Gottlieb, Y., Vaadia, E. and Abeles, M. 1989. Single unit activity in the auditory cortex of a monkey performing a short term memory task. *Experimental Brain Research* 74(1), pp. 139–148.

Greenlee, M.W., Lang, H.J., Mergner, T. and Seeger, W. 1995. Visual short-term memory of stimulus velocity in patients with unilateral posterior brain damage. *The Journal of Neuroscience* 15(3 Pt 2), pp. 2287–2300.

Greenlee, M.W., Rischewski, J., Mergner, T. and Seeger, W. 1993. Delayed pattern discrimination in patients with unilateral temporal lobe damage. *The Journal of Neuroscience* 13(6), pp. 2565–2574.

Griffin, I.C. and Nobre, A.C. 2003. Orienting attention to locations in internal representations. *Journal of Cognitive Neuroscience* 15(8), pp. 1176–1194.

Hansel, D. and Mato, G. 2013. Short-term plasticity explains irregular persistent activity in working memory tasks. *The Journal of Neuroscience* 33(1), pp. 133–149.

Hardman, K.O., Vergauwe, E. and Ricker, T.J. 2017. Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology. Human Perception and Performance* 43(1), pp. 30–54.

Harrison, S.A. and Tong, F. 2009. Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458(7238), pp. 632–635.

Hikosaka, O. and Wurtz, R.H. 1983. Visual and oculomotor functions of monkey substantia nigra pars reticulata. III. Memory-contingent visual and saccade responses. *Journal of Neurophysiology* 49(5), pp. 1268–1284.

Honkanen, R., Rouhinen, S., Wang, S.H., Palva, J.M. and Palva, S. 2015. Gamma Oscillations Underlie the Maintenance of Feature-Specific Information and the Contents of Visual Working Memory. *Cerebral Cortex* 25(10), pp. 3788–3801.

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America* 79(8), pp. 2554–2558.

Howard, M.W., Rizzuto, D.S., Caplan, J.B., Madsen, J.R., Lisman, J., Aschenbrenner-Scheibe, R., Schulze-Bonhage, A. and Kahana, M.J. 2003. Gamma oscillations correlate with working memory load in humans. *Cerebral Cortex* 13(12), pp. 1369–1374.

Inagaki, H.K., Fontolan, L., Romani, S. and Svoboda, K. 2019. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* 566(7743), pp.

212–217.

James, W. 1890. *The Principles of Psychology* . United States: Henry Holt and Company.

Jensen, O. and Tesche, C.D. 2002. Frontal theta activity in humans increases with memory load in a working memory task. *The European Journal of Neuroscience* 15(8), pp. 1395–1399.

Johnson, E.N. and Mullen, K.T. 2016. Color in the cortex. In: Kremers, J., Baraas, R. C., and Marshall, N. J. eds. *Human Color Vision*. Cham: Springer International Publishing, pp. 189–217.

Johnson, J.S., Spencer, J.P., Luck, S.J. and Schöner, G. 2009. A dynamic neural field model of visual working memory and change detection. *Psychological Science* 20(5), pp. 568–577.

Jun, J.K., Miller, P., Hernández, A., Zainos, A., Lemus, L., Brody, C.D. and Romo, R. 2010. Heterogenous population coding of a short-term memory and decision task. *The Journal of Neuroscience* 30(3), pp. 916–929.

Kamiński, J., Sullivan, S., Chung, J.M., Ross, I.B., Mamelak, A.N. and Rutishauser, U. 2017. Persistently active neurons in human medial frontal and medial temporal lobe support working memory. *Nature Neuroscience* 20(4), pp. 590–601.

Kelly, S.P., Lalor, E.C., Reilly, R.B. and Foxe, J.J. 2006. Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention. *Journal of Neurophysiology* 95(6), pp. 3844–3851.

Kepecs, A., Uchida, N., Zariwala, H.A. and Mainen, Z.F. 2008. Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455(7210), pp. 227–231.

van Kerkoerle, T., Self, M.W. and Roelfsema, P.R. 2017. Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nature Communications* 8, p. 13804.

Kilpatrick, Z.P. 2018. Synaptic mechanisms of interference in working memory. *Scientific reports* 8(1), p. 7879.

Kiyonaga, A., Scimeca, J.M., Bliss, D.P. and Whitney, D. 2017. Serial Dependence across Perception, Attention, and Memory. *Trends in Cognitive Sciences* 21(7), pp. 493–497.

Knight, B.W. 1972. Dynamics of encoding in a population of neurons. *The Journal of General Physiology* 59(6), pp. 734–766.

Kornblith, S., Buschman, T.J. and Miller, E.K. 2016. Stimulus load and oscillatory activity in higher cortex. *Cerebral Cortex* 26(9), pp. 3772–3784.

Krishnan, N., Poll, D.B. and Kilpatrick, Z.P. 2018. Synaptic efficacy shapes resource limitations in working memory. *Journal of Computational Neuroscience* 44(3), pp. 273–295.

Kubota, K. and Niki, H. 1971. Prefrontal cortical unit activity and delayed alternation

performance in monkeys. *Journal of Neurophysiology* 34(3), pp. 337–347.

Laming, D. and Laming, J. 1992. F. Hegelmaier: on memory for the length of a line. *Psychological research* 54(4), pp. 233–239.

de Lange, F.P., Heilbron, M. and Kok, P. 2018. How do expectations shape perception? *Trends in Cognitive Sciences* 22(9), pp. 764–779.

Lapicque, L. 2007. Quantitative investigations of electrical nerve excitation treated as polarization. 1907. *Biological Cybernetics* 97(5–6), pp. 341–349.

Leavitt, M.L., Mendoza-Halliday, D. and Martinez-Trujillo, J.C. 2017. Sustained activity encoding working memories: not fully distributed. *Trends in Neurosciences* 40(6), pp. 328–346.

Lester, R.A., Clements, J.D., Westbrook, G.L. and Jahr, C.E. 1990. Channel kinetics determine the time course of NMDA receptor-mediated synaptic currents. *Nature* 346(6284), pp. 565–567.

Li, D., Constantinidis, C. and Murray, J.D. 2018. Testing burst coding models of working memory with spike trains from primate prefrontal cortex. . In: Society for Neuroscience.

Liberman, A., Fischer, J. and Whitney, D. 2014. Serial dependence in the perception of faces. *Current Biology* 24(21), pp. 2569–2574.

Liberman, A., Zhang, K. and Whitney, D. 2016. Serial dependence promotes object stability during occlusion. *Journal of Vision* 16(15), p. 16.

Liebe, S., Hoerzer, G.M., Logothetis, N.K. and Rainer, G. 2012. Theta coupling between V4 and prefrontal cortex predicts visual short-term memory performance. *Nature Neuroscience* 15(3), pp. 456–62, S1.

Lieder, I., Adam, V., Frenkel, O., Jaffe-Dax, S., Sahani, M. and Ahissar, M. 2019. Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nature Neuroscience* 22(2), pp. 256–264.

Lin, P.-H. and Luck, S.J. 2009. The Influence of Similarity on Visual Working Memory Representations. *Visual cognition* 17(3), pp. 356–372.

Luck, S.J. and Vogel, E.K. 1997. The capacity of visual working memory for features and conjunctions. *Nature* 390(6657), pp. 279–281.

Lundqvist, M., Herman, P. and Lansner, A. 2011. Theta and gamma power increases and alpha/beta power decreases with memory load in an attractor network model. *Journal of Cognitive Neuroscience* 23(10), pp. 3008–3020.

Lundqvist, M., Herman, P., Warden, M.R., Brincat, S.L. and Miller, E.K. 2018. Gamma and beta bursts during working memory readout suggest roles in its volitional control. *Nature Communications* 9(1), p. 394.

Lundqvist, M., Rose, J., Herman, P., Brincat, S.L., Buschman, T.J. and Miller, E.K. 2016. Gamma and beta bursts underlie working memory. *Neuron* 90(1), pp. 152–164.

Ma, W.J., Husain, M. and Bays, P.M. 2014. Changing concepts of working memory.

Nature Neuroscience 17(3), pp. 347–356.

Manassi, M., Liberman, A., Chaney, W. and Whitney, D. 2017. The perceived stability of scenes: serial dependence in ensemble representations. *Scientific reports* 7(1), p. 1971.

Manassi, M., Liberman, A., Kosovicheva, A., Zhang, K. and Whitney, D. 2018. Serial dependence in position occurs at the time of perception. *Psychonomic Bulletin & Review* 25(6), pp. 2245–2253.

Markram, H. and Tsodyks, M. 1996. Redistribution of synaptic efficacy between neocortical pyramidal neurons. *Nature* 382(6594), pp. 807–810.

Markram, H., Wang, Y. and Tsodyks, M. 1998. Differential signaling via the same axon of neocortical pyramidal neurons. *Proceedings of the National Academy of Sciences of the United States of America* 95(9), pp. 5323–5328.

Masse, N.Y., Hodnefield, J.M. and Freedman, D.J. 2017. Mnemonic encoding and cortical organization in parietal and prefrontal cortices. *The Journal of Neuroscience* 37(25), pp. 6098–6112.

Matthey, L., Bays, P.M. and Dayan, P. 2015. A probabilistic palimpsest model of visual short-term memory. *PLoS Computational Biology* 11(1), p. e1004003.

Mazaheri, A. and Jensen, O. 2006. Posterior alpha activity is not phase-reset by visual stimuli. *Proceedings of the National Academy of Sciences of the United States of America* 103(8), pp. 2948–2952.

McKeown, D. and Mercer, T. 2012. Short-term forgetting without interference. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 38(4), pp. 1057–1068.

Medendorp, W.P., Kramer, G.F.I., Jensen, O., Oostenveld, R., Schoffelen, J.-M. and Fries, P. 2007. Oscillatory activity in human parietal and occipital cortex shows hemispheric lateralization and memory effects in a delayed double-step saccade task. *Cerebral Cortex* 17(10), pp. 2364–2374.

Miller, G.A. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63(2), pp. 81–97.

Mitchell, D.J., Cusack, R. and Cam-CAN 2018. Visual short-term memory through the lifespan: Preserved benefits of context and metacognition. *Psychology and aging* 33(5), pp. 841–854.

Mongillo, G., Barak, O. and Tsodyks, M. 2008. Synaptic theory of working memory. *Science* 319(5869), pp. 1543–1546.

Moore, T. and Armstrong, K.M. 2003. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* 421(6921), pp. 370–373.

Murphy, S.C., Palmer, L.M., Nyffeler, T., Müri, R.M. and Larkum, M.E. 2016. Transcranial magnetic stimulation (TMS) inhibits cortical dendrites. *eLife* 5.

Murray, J.D., Bernacchia, A., Freedman, D.J., Romo, R., Wallis, J.D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D. and Wang, X.-J. 2014. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience* 17(12),

pp. 1661–1663.

Murray, J.D., Jaramillo, J. and Wang, X.-J. 2017. Working Memory and Decision-Making in a Frontoparietal Circuit Model. *The Journal of Neuroscience* 37(50), pp. 12167–12186.

Nassar, M.R., Helmers, J.C. and Frank, M.J. 2018. Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review* 125(4), pp. 486–511.

Nilsson, T.H. and Nelson, T.M. 1981. Delayed monochromatic hue matches indicate characteristics of visual memory. *Journal of Experimental Psychology. Human Perception and Performance* 7(1), pp. 141–150.

Oberauer, K. and Lin, H.-Y. 2017. An interference model of visual working memory. *Psychological Review* 124(1), pp. 21–59.

Olson, I.R. and Jiang, Y. 2002. Is visual short-term memory object based? Rejection of the “strong-object” hypothesis. *Perception & Psychophysics* 64(7), pp. 1055–1067.

Palva, J.M., Palva, S. and Kaila, K. 2005. Phase synchrony among neuronal oscillations in the human cortex. *The Journal of Neuroscience* 25(15), pp. 3962–3972.

Panichello, M.F., DePasquale, B., Pillow, J.W. and Buschman, T. 2018. Error-correcting dynamics in visual working memory. *BioRxiv*.

Papadimitriou, C., Ferdoash, A. and Snyder, L.H. 2015. Ghosts in the machine: memory interference from the previous trial. *Journal of Neurophysiology* 113(2), pp. 567–577.

Papadimitriou, C., White, R.L. and Snyder, L.H. 2017. Ghosts in the Machine II: Neural Correlates of Memory Interference from the Previous Trial. *Cerebral Cortex* 27(4), pp. 2513–2527.

Parra, M.A., Cubelli, R. and Della Sala, S. 2011. Lack of color integration in visual short-term memory binding. *Memory & Cognition* 39(7), pp. 1187–1197.

Pasternak, T. and Greenlee, M.W. 2005. Working memory in primate sensory systems. *Nature Reviews. Neuroscience* 6(2), pp. 97–107.

Pertzov, Y., Bays, P.M., Joseph, S. and Husain, M. 2013. Rapid forgetting prevented by retrospective attention cues. *Journal of Experimental Psychology. Human Perception and Performance* 39(5), pp. 1224–1231.

Pertzov, Y., Dong, M.Y., Peich, M.-C. and Husain, M. 2012. Forgetting what was where: the fragility of object-location binding. *Plos One* 7(10), p. e48214.

Pertzov, Y., Manohar, S. and Husain, M. 2017. Rapid forgetting results from competition over time between items in visual working memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition* 43(4), pp. 528–536.

Pesaran, B., Pezaris, J.S., Sahani, M., Mitra, P.P. and Andersen, R.A. 2002. Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature Neuroscience* 5(8), pp. 805–811.

Phillips, W.A. and Baddeley, A.D. 1971. Reaction time and short-term visual memory. *Psychonomic science* 22(2), pp. 73–74.

Pina, J.E., Bodner, M. and Ermentrout, B. 2018. Oscillations in working memory and neural binding: A mechanism for multiple memories and their interactions. *PLoS Computational Biology* 14(11), p. e1006517.

Pratte, M.S. 2018. Swap errors in spatial working memory are guesses. *Psychonomic Bulletin & Review*.

Purves, D. 2001. *Neuroscience*. 4th ed. Sunderland, Mass: Sinauer.

Qi, X.-L. and Constantinidis, C. 2012. Variability of prefrontal neuronal discharges before and after training in a working memory task. *Plos One* 7(7), p. e41053.

Rademaker, R.L., Bloem, I.M., De Weerd, P. and Sack, A.T. 2015. The impact of interference on short-term memory for visual orientation. *Journal of Experimental Psychology. Human Perception and Performance* 41(6), pp. 1650–1665.

Reinhart, R.M.G., Heitz, R.P., Purcell, B.A., Weigand, P.K., Schall, J.D. and Woodman, G.F. 2012. Homologous mechanisms of visuospatial working memory maintenance in macaque and human: properties and sources. *The Journal of Neuroscience* 32(22), pp. 7711–7722.

Romero, M.C., Davare, M., Armendariz, M. and Janssen, P. 2018. Neural basis of Transcranial Magnetic Stimulation at the single-cell Level. *BioRxiv*.

Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E. and Postle, B.R. 2016. Reactivation of latent working memories with transcranial magnetic stimulation. *Science* 354(6316), pp. 1136–1139.

Rossi, S., Hallett, M., Rossini, P.M., Pascual-Leone, A. and Safety of TMS Consensus Group 2009. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clinical Neurophysiology* 120(12), pp. 2008–2039.

Roux, F., Wibral, M., Mohr, H.M., Singer, W. and Uhlhaas, P.J. 2012. Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. *The Journal of Neuroscience* 32(36), pp. 12411–12420.

Sajad, A., Sadeh, M., Yan, X., Wang, H. and Crawford, J.D. 2016. Transition from Target to Gaze Coding in Primate Frontal Eye Field during Memory Delay and Memory-Motor Transformation. *eNeuro* 3(2).

Sakai, K., Rowe, J.B. and Passingham, R.E. 2002. Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nature Neuroscience* 5(5), pp. 479–484.

Samaha, J., Switzky, M. and Postle, B.R. 2018. Confidence boosts serial dependence in orientation estimation. *BioRxiv*.

Schneegans, S. and Bays, P. 2018. New perspectives on binding in visual working memory.

Schneegans, S. and Bays, P.M. 2018. Drift in neural population activity causes working memory to deteriorate over time. *The Journal of Neuroscience* 38(21), pp.

4859–4869.

Schneegans, S. and Bays, P.M. 2017a. Neural architecture for feature binding in visual working memory. *The Journal of Neuroscience* 37(14), pp. 3913–3925.

Schneegans, S. and Bays, P.M. 2017b. Restoration of fMRI Decodability Does Not Imply Latent Working Memory States. *Journal of Cognitive Neuroscience* 29(12), pp. 1977–1994.

Schoenfeld, M.A., Hopf, J.-M., Merkel, C., Heinze, H.-J. and Hillyard, S.A. 2014. Object-based attention involves the sequential activation of feature-specific cortical modules. *Nature Neuroscience* 17(4), pp. 619–624.

Scimeca, J.M., Kiyonaga, A. and D'Esposito, M. 2018. Reaffirming the sensory recruitment account of working memory. *Trends in Cognitive Sciences* 22(3), pp. 190–192.

Shadlen, M.N. and Movshon, J.A. 1999. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. *Neuron* 24(1), pp. 67–77, 111.

Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T., Henson, R.N., Brayne, C., Matthews, F.E. and Cam-CAN 2014. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology* 14, p. 204.

Shin, H., Zou, Q. and Ma, W.J. 2017. The effects of delay duration on visual working memory for orientation. *Journal of Vision* 17(14), p. 10.

Shipstead, Z., Harrison, T.L. and Engle, R.W. 2016. Working memory capacity and fluid intelligence: maintenance and disengagement. *Perspectives on psychological science : a journal of the Association for Psychological Science* 11(6), pp. 771–799.

Siegel, M., Warden, M.R. and Miller, E.K. 2009. Phase-dependent neuronal coding of objects in short-term memory. *Proceedings of the National Academy of Sciences of the United States of America* 106(50), pp. 21341–21346.

Singer, W. 1999. Neuronal synchrony: a versatile code for the definition of relations? *Neuron* 24(1), pp. 49–65, 111.

Softky, W.R. and Koch, C. 1993. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *The Journal of Neuroscience* 13(1), pp. 334–350.

Sotiropoulos, G., Seitz, A.R. and Seriès, P. 2011. Changing expectations about speed alters perceived motion direction. *Current Biology* 21(21), pp. R883–R884.

Souza, A.S., Rerko, L., Lin, H.-Y. and Oberauer, K. 2014. Focused attention improves working memory: implications for flexible-resource and discrete-capacity models. *Attention, perception & psychophysics* 76(7), pp. 2080–2102.

Spaak, E., Watanabe, K., Funahashi, S. and Stokes, M.G. 2017. Stable and dynamic coding for working memory in primate prefrontal cortex. *The Journal of Neuroscience* 37(27), pp. 6503–6516.

Sprague, T.C., Ester, E.F. and Serences, J.T. 2016. Restoring latent visual working memory representations in human cortex. *Neuron* 91(3), pp. 694–707.

Sreenivasan, K.K., Vytlacil, J. and D'Esposito, M. 2014. Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. *Journal of Cognitive Neuroscience* 26(5), pp. 1141–1153.

Standage, D. and Paré, M. 2018. Slot-like capacity and resource-like coding in a neural model of multiple-item working memory. *Journal of Neurophysiology* 120(4), pp. 1945–1961.

Sterratt, D., Graham, B., Gillies, A. and Willshaw, D. 2011. *Principles of computational modelling in neuroscience*. Cambridge: Cambridge University Press.

Stevens, C.F. and Wang, Y. 1995. Facilitation and depression at single central synapses. *Neuron* 14(4), pp. 795–802.

Stokes, M.G. 2015. "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences* 19(7), pp. 394–405.

Suárez-Pinilla, M., Seth, A.K. and Roseboom, W. 2018. Serial dependence in the perception of visual variance. *Journal of Vision* 18(7), p. 4.

Sugase-Miyamoto, Y., Liu, Z., Wiener, M.C., Optican, L.M. and Richmond, B.J. 2008. Short-term memory trace in rapidly adapting synapses of inferior temporal cortex. *PLoS Computational Biology* 4(5), p. e1000073.

Supèr, H., Spekreijse, H. and Lamme, V.A. 2001. A neural correlate of working memory in the monkey primary visual cortex. *Science* 293(5527), pp. 120–124.

Suzuki, M. and Gottlieb, J. 2013. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience* 16(1), pp. 98–104.

Taubert, J., Alais, D. and Burr, D. 2016. Different coding strategies for the perception of stable and changeable facial attributes. *Scientific reports* 6, p. 32239.

Taubert, J., Van der Burg, E. and Alais, D. 2016. Love at second sight: Sequential dependence of facial attractiveness in an on-line dating paradigm. *Scientific reports* 6, p. 22740.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-Can and Henson, R.N. 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* 144(Pt B), pp. 262–269.

Tegnér, J., Compte, A. and Wang, X.-J. 2002. The dynamical stability of reverberatory neural circuits. *Biological Cybernetics* 87(5–6), pp. 471–481.

Todd, J.J. and Marois, R. 2004. Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature* 428(6984), pp. 751–754.

Tsodyks, M., Pawelzik, K. and Markram, H. 1998. Neural networks with dynamic synapses. *Neural Computation* 10(4), pp. 821–835.

Tuckell, H.C. 1988. *Introduction to Theoretical Neurobiology: Volume 2, Nonlinear*

and Stochastic Theories. Cambridge University Press, 1988 ed.

Vogel, E.K. and Machizawa, M.G. 2004. Neural activity predicts individual differences in visual working memory capacity. *Nature* 428(6984), pp. 748–751.

Wandell, B.A., Dumoulin, S.O. and Brewer, A.A. 2007. Visual field maps in human cortex. *Neuron* 56(2), pp. 366–383.

Wang, X.J. 1999. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *The Journal of Neuroscience* 19(21), pp. 9587–9603.

Wang, X.J. 2001. Synaptic reverberation underlying mnemonic persistent activity. *Trends in Neurosciences* 24(8), pp. 455–463.

Wang, Y., Markram, H., Goodman, P.H., Berger, T.K., Ma, J. and Goldman-Rakic, P.S. 2006. Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nature Neuroscience* 9(4), pp. 534–542.

Warren, J.M., Leary, R.W., Harlow, H.F. and French, G.M. 1957. Function of association cortex in monkeys. *The British Journal of Animal Behaviour* 5(4), pp. 131–138.

Watanabe, K. and Funahashi, S. 2014. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nature Neuroscience* 17(4), pp. 601–611.

Wei, Z., Wang, X.-J. and Wang, D.-H. 2012. From distributed resources to limited slots in multiple-item working memory: a spiking network model with normalization. *The Journal of Neuroscience* 32(33), pp. 11228–11240.

Wheeler, M.E. and Treisman, A.M. 2002. Binding in short-term visual memory. *Journal of Experimental Psychology. General* 131(1), pp. 48–64.

Wilken, P. and Ma, W.J. 2004. A detection theory account of change detection. *Journal of Vision* 4(12), pp. 1120–1135.

Wimmer, K., Compte, A., Roxin, A., Peixoto, D., Renart, A. and de la Rocha, J. 2015. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nature Communications* 6, p. 6177.

Wimmer, K., Nykamp, D.Q., Constantinidis, C. and Compte, A. 2014. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience* 17(3), pp. 431–439.

Wolfe, J.M. 1994. Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review* 1(2), pp. 202–238.

Wolff, M.J., Ding, J., Myers, N.E. and Stokes, M.G. 2015. Revealing hidden states in visual working memory using electroencephalography. *Frontiers in Systems Neuroscience* 9, p. 123.

Wolff, M.J., Jochim, J., Akyürek, E.G. and Stokes, M.G. 2017. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience* 20(6), pp. 864–871.

Worden, M.S., Foxe, J.J., Wang, N. and Simpson, G.V. 2000. Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *The Journal of Neuroscience* 20(6), p. RC63.

Xia, Y., Liberman, A., Yamanashi Leib, A. and Whitney, D. 2015. Serial Dependence in the perception of attractiveness. *Journal of Vision* 15(12), p. 1219.

Xu, Y. 2002. Limitations of object-based feature encoding in visual short-term memory. *Journal of Experimental Psychology. Human Perception and Performance* 28(2), pp. 458–468.

Xu, Y. 2017. Reevaluating the sensory account of visual working memory storage. *Trends in Cognitive Sciences* 21(10), pp. 794–815.

Xu, Y. 2018. Sensory cortex is nonessential in working memory storage. *Trends in Cognitive Sciences* 22(3), pp. 192–193.

Zaksas, D. and Pasternak, T. 2006. Directional signals in the prefrontal cortex and in area MT during a working memory for visual motion task. *The Journal of Neuroscience* 26(45), pp. 11726–11742.

Zhang, W. and Luck, S.J. 2008. Discrete fixed-resolution representations in visual working memory. *Nature* 453(7192), pp. 233–235.

Zhang, W. and Luck, S.J. 2009. Sudden death and gradual decay in visual working memory. *Psychological Science* 20(4), pp. 423–428.

Zucker, R.S. and Regehr, W.G. 2002. Short-term synaptic plasticity. *Annual Review of Physiology* 64, pp. 355–405.